

MATHEMATICS OF STATISTICS

PART ONE

BY

JOHN F. KENNEY

University of Wisconsin

SECOND EDITION, SECOND PRINTING

SCIENCE BOOK AGENCY
P-133B, LAKE TERRACE
CALCUTTA-29

This book or any part thereof may not be reproduced in any form without written permission from the author and the publisher

First Published, May 1939

Reprinted January 1941, February 1942

July 1943, March 1944, June 1945

Second Edition, January 1947

Reprinted May 1947

Printers :

**Pradip Kr. Banerjee, Manasi Press,
73 Manicktollé St.
Calcutta.**



Matri Meae

in Signum

Gratitudinis



PREFACE TO THE SECOND EDITION

Striking examples of the utility and scope of the science of statistics have occurred in recent years. As Professor Harold Hotelling remarks (*Annals of Mathematical Statistics*, vol. 11, 1940, pp. 457-470):

Indeed it seems as if the exploitation of the business and manufacturing possibilities of statistical methods has only begun and that limitless further fields are coming into view.

The widespread use of statistical methods and the gratifying interest shown in the present book have made possible a second edition at this time. This opportunity has been used to polish and clarify certain portions of the text. For suggestions leading to the excision of obscurities I am indebted to many of my students and to a number of friends in other universities, particularly to Professors Irving W. Burr, John H. Curtiss, Henry Scheffé, Guy G. Specker, and Howard E. Wahlert. Of course, full responsibility for any remaining errors or other defects is my own.

J. F. K.

PREFACE TO THE FIRST EDITION

The field of statistics is many sided and ranges over different levels. However, between the levels of clerical work at one extreme and mathematical research at the other extreme, there is a well-defined methodology, mathematical in nature, which underlies the specialized applications in the departments of economics, psychology, education, and biology.

This book is an elementary text dealing with the mathematics of statistics. Fortunately, a considerable part of the descriptive methodology of statistics can be understood by those having relatively little knowledge of college mathematics. Although no mathematics beyond the ordinary Freshman course in college algebra is required for a profitable reading of this text, a certain degree of mathematical maturity and intelligence is presupposed. To achieve the maximum success perhaps only the best of those students whose mathematical preparation is limited to the minimum prerequisite should be encouraged to study it. Occasionally, material is introduced to sharpen the interest and challenge the ability of the more advanced student without interrupting the main developments or discouraging those less mature.

In writing this book, considerable selection of material necessarily had to be made. The omission of certain topics will be noted in the table of contents. Judging from my own experience, and that of others, the theory of sampling cannot be taught satisfactorily at the level for which Part I is intended. At best only a superficial use of formulas could be hoped for. Consequently, I have elected to defer this subject to Part II where a systematic treatment can be given. With regard to time series analysis, Professor J. Neyman says in his *Lectures And Conferences On Mathematical Statistics* (p. 106),

We start by trying to split each of the series into several parts, which we arbitrarily assume to be additive. One of these parts is the trend, which we estimate perhaps by fitting a low order parabola to the whole series available. The next part is the "business cycle." The third part is the "seasonal variation," which we frequently estimate by calculating moving averages. Finally, the remainder is considered to arise from random causes, and we concentrate on the question whether such a remainder in one of the variables is correlated

with that in some other. All this procedure seems to me very artificial and arbitrary. . . . In my opinion the whole problem of time series must be treated from a point of view that is quite different from the traditional one just described.

I concur in this opinion and I believe that no useful purpose would be served by drilling students in the traditional procedures.

Throughout the book the student is encouraged and stimulated to master fundamental principles and concepts. Essentially, the job of every statistician is to take hold of situations and disentangle them by the techniques of the science. Therefore, considerable emphasis is placed on technique. I have tried to develop in the student the ability to use symbolism creatively as a language. Numerous examples are given to clarify concepts and illustrate processes. Over two hundred exercises are included. It is intended that these exercises should be handled as in a mathematics course. No laboratory, so-called, is necessary.

Nowadays, no little importance is attached to motivation. I have constantly held in mind the necessity of making the subject interesting and stimulating to the beginning student. Nevertheless, I venture the opinion that the best motivation for intelligent students is the feeling that their teacher knows his subject.

In preparing the manuscript a large number of books and papers have been examined and perhaps leaned upon. No claim to originality is made except possibly in the matter of arrangement and pedagogical approach. Numerous references to the scholarly achievements of others are cited. It is hoped that the serious student will read some of these and thereby widen his perspective and enhance his interest.

In conclusion, I wish to express my deep appreciation to Professor Allen T. Craig and Dr. Mason F. Wescott who critically read the manuscript and made many suggestions for its improvement.

April, 1939

JOHN F. KENNEY

CONTENTS

INTRODUCTION

SECTION	PAGE
1. Definition	1
2. Scope	1
3. Statistical Methods in the Social Sciences	2
4. Mathematics and Statistics	3
5. Problem Assignments	4
6. Calculating Machines	4
7. Collateral Reading	5

CHAPTER I

FREQUENCY DISTRIBUTIONS

1. Variables and Constants	7
✓ 2. Variates	7
3. Accuracy of Measurements	8
4. Necessity for Classification	9
5. Tabulation	9
✓ 6. Frequency Distribution	10
7. Class Intervals	12
8. Distinction between Class Limits and Class Boundaries	15
9. Rules for Making a Frequency Distribution	16
✓ 10. Cumulative Frequencies	16
11. Additional Distributions	18

CHAPTER II

GRAPHICAL REPRESENTATION

1. The Function Concept	22
2. Charts	24
3. Frequency Polygon	24
4. Histogram	25
5. Frequency Curves	25
6. Ogives	27
7. Relation of <i>Cum f</i> to Areas	28

CHAPTER III

AVERAGES

1. Introductory	29
2. Notation	29
3. Arithmetic Mean	33

SECTION	PAGE
4. Weighted Arithmetic Mean	33
5. Arithmetic Mean from Frequency Table	35
6. Translation of Axes; Deviations	36
7. Properties of \bar{x}	38
8. Short Methods of Computing \bar{x}	39
9. Geometric Explanation	41
10. Mean of Means	43
11. The Mode	47
12. The Median	47
13. Median of a Frequency Distribution	47
14. Graphical Interpretation of Mean, Median, and Mode	50
15. Discussion	51
16. Geometric Mean	52
17. Harmonic Mean	55

CHAPTER IV

MOMENTS

1. Moments about an Arbitrary Origin	62
2. Moments in Units of the Class Interval	64
3. Moments about the Mean	64
4. Relations between the μ 's and the ν 's	65
5. Standard Deviation	68
6. Standard Units	69
7. Moments in Standard Units	72
8. Use of α_3 and α_4	73
9. Summary	74
10. Sheppard's Corrections	78

CHAPTER V

MEASURES OF DISPERSION

1. Introduction	81
2. The Quartile Deviation	82
3. Mean Deviation	84
4. The Standard Deviation	86
5. Relative Dispersions	90
6. Scaling a Distribution in Terms of σ	90
7. Semi-Interquartile Range in Terms of σ	92
8. N Small, Ungrouped Data	93
9. Standard Deviation of Combinations of Sets	99
10. Graphical Representation	103

CHAPTER VI

TYPES OF DISTRIBUTIONS. THE NORMAL CURVE

1. Skewness and Kurtosis	109
2. Frequency Curves	112

Contents

xi

SECTION	PAGE
3. The Normal Curve.....	114
4. Standard Form.....	115
5. Tables of Standard Ordinates and Areas.....	116
6. Properties.....	118
7. Curve Fitting.....	124
8. Graduation.....	128
9. Purpose of a Graduation.....	128
10. Probability.....	131
11. Probability Paper.....	132

CHAPTER VII CURVE FITTING

1. Empirical Expressions.....	136
2. Linear Functions.....	137
3. Quadratic Function.....	138
4. Fitting a Straight Line.....	140
5. Graphically.....	140
6. Method of Moments.....	141
7. An Alternative Procedure.....	144
8. Least Squares.....	144
9. Simplification.....	149
10. Time Series.....	150
11. Exponential Trends.....	152
12. Further Remarks on the Exponential Function.....	156
13. Ratio Charts.....	157
14. Logarithmic Coordinate Paper.....	161
15. Parabolic Trend.....	162
16. The Gompertz Curve.....	164
17. Remarks and References.....	166

CHAPTER VIII CORRELATION THEORY

1. The Meaning of Simple Correlation.....	170
2. The Coefficient of Correlation.....	171
3. Other Formulas for r	173
4. Regression.....	177
5. The Standard Error of Estimate.....	179
6. Properties of the Correlation Coefficient and Standard Error.....	180
7. Further Discussion.....	183
8. Coefficient of Alienation.....	185
9. Correlation Table.....	189
10. Notation.....	190
11. Means and Variances.....	192
12. Computation of Means.....	194
13. Computation of r	195
14. Remarks on Computation of r	199

SECTION	PAGE
15. Regression Lines for a Correlation Table	199
16. Applications	202
17. S_y for a Correlation Table	205
18. Normal Correlation Surface	206
19. Properties of Normal Bivariate Surface	208
20. Reliability of Predictions	209
21. Non-Linear Regression. Correlation Ratio	212
22. Computation of η^2	214
23. Further Discussion. Test for Linearity of Regression	217
24. Correlation from Ranks	222
25. Interpretation. Common Elements	225
Review Questions and Problems	227
Tables	235
Index	259

MATHEMATICS OF STATISTICS

INTRODUCTION

1. Definition. The word *statistics* is used in at least two different senses. Construed as plural it refers to the systematic presentation of quantitative data. Used in a singular sense, the word *statistics* refers to the science which has for its object the classification and analysis of quantitative data so that intelligent judgments may be passed upon them.

It is usually clear from the context which meaning¹ is intended, although some persons prefer the expression *statistical methods* for this second meaning. Statistical methods are all those devices used in the collection and analysis of data. The *theory of statistics* is the exposition of statistical methods and is of a mathematical nature.

2. Scope. There used to be a widespread misapprehension that statistics is a branch of economics. As a matter of fact, statistical problems arise in many different fields — biology, economics, engineering, insurance, education, physics, and astronomy, as well as various branches of business. The exploration of certain aspects of nearly every field involves some phase of statistical theory. Indeed, certain types of statistical methodology may have almost unexpected applications — the discovery, for example, that the life of physical property² is governed by much the same statistical rules as govern the lives of human beings, and hence, that life tables may be applied to both. Physicists have discovered that many of the problems in the modern theory of the structure of the atom are essentially statistical in nature. In recent years industrial companies have placed an increasing reliance on statistical methods in controlling the quality of goods during manufacture.

Statistics as a science is making contributions to all the sciences. On the other hand, some sciences like biometry and physics have

¹ In addition to the two meanings given above, another has crept into the recent literature where reference is made to a *statistic*. This term will be explained later.

² *Life Expectancy of Physical Property* — E. B. Kurtz. Ronald Press.

contributed much in the development of statistics and its terminology. The following quotation from *Science* may appropriately be mentioned here:

The extension of the scope of quantitative methods through the medium of statistical analysis is one of the most significant things going on in the scientific world at the present time.¹

The importance of statistical method in present-day thinking has been well stated, as follows:

More and more the modern temper relies upon statistical method in its attempts to understand and to chart the workings of the world in which we live. Particularly in those sciences which deal with human beings, whether in their physical and biological aspects or in their social, economic, and psychological relations, the spirit of our time asks that its conclusion be based not so much upon the distinctive reactions of one or two individuals as upon the observation of large numbers of individuals, the measurement of their common likenesses and the extent of their diversity. As the data thus gathered from mass phenomena become extensive, it becomes imperative to have methods of organization to bring the facts within the compass of our understanding, methods of analysis to make the essential relations appear out of the mass of detail in which they are hidden, and methods of classification and description to facilitate the presentation of the data for the study and consideration of other persons. Thus statistical method becomes a telescope through which we can study a larger terrain than would be accessible to our unaided vision.²

3. Statistical Methods in the Social Sciences. Because statistics is fundamentally the study of aggregates of individuals, rather than of individuals, whether these *individuals* be observations or measurements or persons, it is apparent that statistical methods are essential to social studies. Indeed it has been said that it is principally by the aid of such methods that these studies may be raised to the rank of sciences.

This particular dependence of social studies upon statistical methods is mentioned in a recent book³ from which we quote the following:

If, as seems probable, our present uncoördinated large-scale business is to be further developed into an efficiently managed instrument of production serving the needs of the people, then statistics, together with mathematical economics, will emerge among the most important tools of the social sciences. For it is by

¹ *Science*. January 18, 1929.

² *Mathematics and Statistics* — Walker. Sixth Yearbook, National Council of Teachers of Mathematics.

³ Reprinted by permission from *Methods of Statistical Analysis* by Davies and Crowder, published by John Wiley and Sons, Inc.

means of averages, dispersions, coefficients of variability, trends, and regressions, as pictured in control charts, that management is able to visualize and direct the movements of large masses of population.

The work of the statistician is much like that of the map maker who presents the traveler with a sketch of important highways, showing the locations of towns and geographical features. The map is not a picture of reality. It shows cities as dots, and rivers as lines. It has purposely omitted the interesting details of scenery and the still more important features of human interest which lie along the route and which constitute the traveler's real objectives. Nevertheless, as a means of reaching these objectives, the map is extremely useful. And so it is with statistics in the hands of the business executive and statesman. Back of the charts are human beings with their varying characteristics and vital interests, few of which can be described in figures. Yet as a means of serving these interests, of keeping trade moving from one region to another, of allocating investment and labor, and of apportioning relief to maladjusted industries and dependent classes, statistics and mathematical methods are important, and are becoming increasingly important with the growing complexity of society.

It may be said that the study of statistics is not merely an attempt to describe what actually occurs, though it must begin at this point, but in its broader aspects it is the logical background of business and social management. Hence what appears now to be mere abstraction may later become the basic necessity of an applied science. Eventually, it may be assumed, the social arts of business and politics will rest upon as substantial a theoretical and mathematical background as physics, chemistry, and engineering.

4. Mathematics and Statistics. Statistical problems are of interest, therefore, not only to the worker in the particular field but also to the mathematician, inasmuch as methods adequate to the treatment of these problems can best be presented in the precise and accurate language of mathematics. Moreover, statistical methods are grounded in statistical theory which is a branch of applied mathematics.

Although it is true that some statistical problems are ultimately problems in advanced mathematics, many of which mathematicians have not yet been able to solve, nevertheless a large and interesting part of statistical analysis requires mathematics no more advanced than elementary algebra.

It has been said that sooner or later every true science tends to become mathematical. The notation of mathematics is simply a language and it is not limited to any particular field of knowledge. The following quotations are inserted to help the student approach the study of statistics in the proper spirit.

1. **Mathematics, the science of the ideal, becomes the means of investigating, understanding, and making known the world of the real.** — White.

2. Probably among all the pursuits of the university, mathematics preeminently demands self-denial, patience, and perseverance. . . . — Todhunter.

3. From time immemorial, there has been but one way to become a mathematician and there will never be another: it is a way interior to the subject and involves years of assiduous toil. Short-cuts to mathematical scholarship there are none, whether the seeker be a philosopher or a king. — Keyser.

4. Will is the creative force. Without the will to learn there is no learning. And when the will is feeble and confused, learning lags. — Mursell.

5. The theory of statistics is not easy, not so much because it is abstruse, as because the ideas are new to most people, and a good deal of hard thinking and patient work will be necessary. . . . Statistical work always involves a lot of computing [and] there is no better way of learning statistics than by working through examples. — Tippet.

5. Problem Assignments. The student should realize at the outset that statistical methods are not substitutes for thinking but are aids and supplements to it. A superficial knowledge of statistical technique cannot take the place of good judgment. Mere ability to substitute in formulas should not be confused with genuine statistical sophistication and insight. To the serious and capable student who intends to master this course, formulas will be a set of functioning concepts and tools rather than machines into which material may be fed to grind out a meaningless answer.

This opportunity is also taken to point out that even mathematical discourse consists of sentences. Punctuation should not be omitted in sequences of equations and other mathematical statements. (It is admitted, however, that many of us find this difficult to remember.)

Throughout the book exercises are inserted to give the student an opportunity to test his knowledge of the theory and methodology, and to develop his power of analysis. In grading the solutions, value will be attached to accuracy, thoroughness, neatness, and systematic arrangement of the work.

6. Calculating Machines.¹ A full description of the parts of a calculating machine and their operation may be obtained from an *Instruction Book* which is furnished by the manufacturer, so only a brief description will be given here.

A calculating machine is constructed to add and subtract. By means of continued addition or subtraction, operations involving multiplication, division, and square root can also be performed with great speed.

¹ The early history of modern computing machines is outlined in the *American Mathematical Monthly*, vol. 31 (1924), pp. 422–429.

In addition to a keyboard on which numbers can be punched, most machines have a sliding carriage, carrying two dials one above the other. These dials are called *revolution register* (upper dial) and *product register* (lower dial). In finding a product nx , one of the factors n is punched on the keyboard and as the motive crank at the side is turned,¹ the other factor x appears on the upper dial. The product nx is then read from the lower dial.

An important property of the modern calculating machine is its adaptability to short cuts and combinations of operations. For example, one may multiply two numbers nx together and add the result to a third number k without tabulating the intermediate steps. This is accomplished by punching the number k on the keyboard, transferring it to the lower dial (product register), and then proceeding as in finding the product nx . The result $nx + k$ is then read from the lower dial. An extension of this procedure is especially useful in a series of computations where k and n are constant and various values are assigned to x . To describe the procedure, suppose it is required to calculate the successive values of $12 + 6x$ for $x = 5, 7, 15, 12$, etc. The number $k = 12$ is first registered on the lower dial, then the factor $n = 6$ is placed on the keyboard, and by turning the crank forward five times to make the first value of $x = 5$ appear on the upper dial, the result $12 + 6 \times 5$ appears on the lower dial. Instead of clearing the dial, the crank is now turned forward twice more to rebuild the value $x = 5$ into $x = 7$, and the result $12 + 6 \times 7$ can be read from the lower dial. In rebuilding $x = 15$ into $x = 12$ the crank is turned backwards. This procedure can be repeated until all the required values of $12 + 6x$ have been calculated. A process of this sort is called the *continuous method* of calculating.

In most of the exercises in this course, the computations are not laborious and calculating machines are not required. However, if machines are available they may be used to advantage in Chapters IV and VI. The student who desires to develop skill on a calculating machine should begin now to study an *Instruction Book* and practice the fundamental operations explained there.

7. Collateral Reading. Perhaps no single textbook can meet all the needs of all students of statistics. There are several good books on elementary statistics which, although not fundamentally different,

¹ The beginner will probably wish to practice on a manually operated machine before attempting to use the high-speed electric and automatic machines.

present different points of view on certain topics and treat them with varying degrees of emphasis depending upon the field of major interest. At least some of the books listed below should be readily available on the reserve shelf of the library. The list should be useful to those who wish to study more fully certain details in which they may be interested.

1. Bivins — *The Ratio Chart in Business*. Codex Book Co.
2. Burgess — *The Mathematics of Statistics*. Houghton Mifflin and Co.
3. Camp — *The Mathematical Part of Elementary Statistics*. D. C. Heath and Co.
4. Deming — *Statistical Adjustment of Data*. John Wiley & Sons, Inc.
5. Freeman — *Industrial Statistics*. Wiley.
6. Garrett — *Statistics in Psychology and Education*. Longmans, Green and Co.
7. Glover — *Tables of Applied Mathematics*. Wahr.
8. Haskell — *Graphic Charts in Business*. Codex Book Co.
9. Mills — *Statistical Methods, Revised*. Henry Holt and Co.
10. Pearl — *Medical Biometry and Statistics*. W. B. Saunders and Co.
11. Rider — *Statistical Methods*. Wiley.
12. Scarborough — *Numerical Mathematical Analysis*. The Johns Hopkins Press.
13. Snedecor — *Statistical Methods*. Collegiate Press, Inc., Ames, Iowa.
14. Treloar — *Statistical Reasoning*. Wiley.
15. Walker — *Elementary Statistical Methods*. Holt.
16. Yule and Kendall — *The Theory of Statistics*. Griffin and Co.

CHAPTER I

FREQUENCY DISTRIBUTIONS

1. Variables and Constants. A *variable* is a number symbol which may take on any value in a set of values which is called its *range*. A *constant* is a symbol whose range consists of only one value (in a particular discussion or situation). Letters toward the end of the alphabet, such as x , y , u , and v , are commonly used to denote variables. When a constant does not have a definite value such as 2, $\frac{1}{3}$, π , and so forth, it is customary to represent the constant by a letter toward the beginning of the alphabet.

Two famous constants are

$$\pi = 3.14159 \dots, \quad e = 2.71828 \dots$$

They occur in mathematics in many important, interesting, and even curious ways. As instances of the latter, the following examples are noteworthy.

$$e = 2 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \dots, \quad \text{where } n! = n(n-1)(n-2) \dots 1.$$

$$\frac{\pi}{4} = \frac{1}{1 + \frac{1^2}{2 + \frac{3^2}{2 + \frac{5^2}{2 + \frac{7^2}{2 + \dots}}}}}$$

The expression for e is called a convergent infinite series and that for $\pi/4$ a continued fraction.

2. Variates. In general, statistical data are obtained by taking observations or measurements on one or more variables. The values thus obtained are sometimes called *variates*.¹ For example, in computing the average monthly rainfall of a region the variable is rainfall and the amount of rainfall for any month is a variate. Like-

¹ A somewhat different usage of this term is explained in Part II.

wise, if the bank clearings of the city of Madison are under consideration, then the variable is bank clearings, and the clearings for any specified interval are variates. If we denote a variable by x then the N values which it takes on are denoted by x_1, x_2, \dots, x_N .

Variates are of two kinds: *continuous* and *discrete*. Continuous variates are values of a variable which, theoretically, can be measured to any degree of fineness, such as heights, weights, temperatures, ages. All the numbers between $x = 0$ and $x = 1$ form a set of continuous variates. But if we restrict x to the rational numbers in this interval we have a set of separate and distinct values with "vacant" spaces between them. Values of a variable which are thus restricted to particular values in order to have any meaning are called discrete variates. Other examples of discrete variates are: size of families, closing prices of stocks, "successes" in tossing a coin. A set of discrete variates is usually obtained by counting whereas continuous variates are usually obtained by measurement.

3. Accuracy of Measurements. In the case of continuous variates, the observed values as recorded can never be absolutely established by measurement. Thus, the height or weight of an object can be measured only approximately, the error depending upon the precision of the instrument and the care and accuracy of the observer. However, it is not always necessary that measurements be recorded as accurately as it is possible to make them. Similarly, in the case of discrete variates the standard of accuracy used may be less than it is possible to obtain. In population statistics, for example, it may be sufficient to record the numbers to the nearest thousand, with three zeros at the end to fill out to the decimal point. Thus,

<i>City</i>	<i>Population</i>
A	326,000
B	729,000

On the other hand, the exact number of students in a university might be required. The degree of accuracy needed is determined by the purpose of the investigation and it is limited by the closeness with which the variables can be measured.

It follows, therefore, that the degree of accuracy in the final result of a problem involving computations is limited by that of the original data. Students sometimes carry results of problems to five or more decimal places when the original data do not justify more than two

or three decimal places. A table of measurements which constitutes the raw data for a statistical investigation should always specify the degree of accuracy in the readings. Thus, if monthly rainfall is being measured to the nearest hundredth of an inch, and one measurement seems to be exactly 5 inches, it should be recorded as 5.00 inches, with two zeros. A measurement that is merely recorded as 5 means it is correct to the nearest integer and its true value lies between 4.5 and 5.5, whereas 5.00 means the true value is known to lie between 4.995 and 5.005. The three digits in 5.00 are said to be significant.

4. Necessity for Classification. After the data have been collected in any statistical investigation the first step has to do with introducing order in the raw material. Usually we have some hundreds of variates which have been recorded merely in the arbitrary order in which the observations or measurements happened to be made. But in order to analyze a series of variates so that intelligent judgments may be formed about it or that comparisons may be made between two series of variates, proper classification is necessary and of prime importance.

Such classification is not always an easy thing to effect, because it is the one part of statistical methods for which no very definite rules can be given. Most people, until they have tried, imagine that to collect and arrange data in classes and in tables is a straightforward procedure involving no great technique or experience. Although much can be learned from a careful study of the illustrations and discussions that appear in the following pages and the compilations of reputable bureaus such as the census volumes, nevertheless, experience is the best teacher in effecting the most appropriate classification for any set of variates.

5. Tabulation. In carrying out the process of classification, it becomes natural to arrange the results in tabular form, setting forth clearly and explicitly the statistics one wishes to present. In drawing up any table the following general rules should be observed:

- (1) Every table must be self-explanatory. To accomplish this the title should be short, but not at the expense of clearness.
- (2) Full explanatory notes, when necessary, should be incorporated in the table, either directly under the descriptive title and before the body of the table, or else directly under the form.
- (3) The columns and rows should be arranged in a logical order to facilitate comparisons.

- (4) In tabulating long columns of figures, spaces should be left after every five or ten rows. Long unbroken columns are confusing, especially when one is comparing two numbers in a row but in widely separated columns.
- (5) If the numbers tabulated have more than three significant figures, the digits should be grouped in threes. Thus, one should write 4 685 732, not 4685732.
- (6) Double lines at the top (or at the top and bottom) may enhance the effectiveness of a table. If the table nicely fills the width of the page, no side lines should be used. In such cases the omission of the side lines will have the tendency to emphasize the other vertical lines and cause the interior columns to stand out better. The columns should not be widely separated and the form of a narrow, compact table should have its side lines.

The following points are particularly important in practical work:

- (7) Source of data should be included.
- (8) Units of the data presented should be clear.
- (9) Accuracy of transcription must not only be striven for but actually achieved. A reader who finds one error (even though this be the only one) is likely to disparage the whole table.

TABLE 1 — GRADES OF 100 STUDENTS IN FRESHMAN MATHEMATICS

75	86	66	86	50	78	66	79	68	60
80	83	87	79	80	77	81	92	57	52
58	82	73	95	66	60	84	80	79	63
80	88	58	84	96	87	72	65	79	80
86	68	76	41	80	40	63	90	83	94
76	66	74	76	68	82	59	75	35	34
65	63	85	87	79	77	76	74	76	78
75	60	96	74	73	87	52	98	88	64
76	69	60	74	72	76	57	64	67	58
72	80	72	56	73	82	78	45	75	56

6. Frequency Distribution. From the standpoint of a mathematical analysis of statistics, the most important form of tabulation is the so-called frequency distribution. Rough data do not present any clear ideas of description unless they are organized and condensed in a systematic way. We therefore partition the raw data into *classes* of appropriate size, showing the corresponding frequency of *variates* in each class. When any set of statistics is systematically

arranged in this way it is called a frequency distribution. For example, upon an examination of the raw data of Table 1, it is difficult to state any very definite conclusions as to whether these grades represent preponderantly good students or poor ones. The frequency distribution of Table 2, however, does give us more precise infor-

TABLE 2 — FREQUENCY TABLE OF 100 GRADES

<i>Class Limits</i>	<i>Tally Marks</i>	<i>Frequency</i>
30-39	//	2
40-49	///	3
50-59	+++ +++ /	
60-69	+++ +++ +++ +++	
70-79	+++ +++ +++ +++ +++ +++ //	
80-89	+++ +++ +++ +++ +++	25
90-99	+++ //	7
<i>Total</i>		100

mation. We see at a glance that there were 32 students with grades between 70 and 80, and that all but 16 had grades of 60 or above. In Table 3, the confusion of detail is still more apparent. The corresponding frequency distribution is given in Table 4.

The width of a class is called the class interval, and in general the successive class intervals should be of equal width. The mid-value of such an interval is variously called the class mark, mid-value, central value. The width of a class interval is therefore seen to be the common difference between two consecutive class marks. It is also the difference between the lower (or upper) limit of two successive classes. Thus, in Table 4, the class interval is half an inch and the successive class marks are 0.245, 0.745, etc., inches.

7. Class Intervals. Grouping variates into the most appropriate number of classes is a matter of judgment. The choice of intervals to be used in tabulating any particular set of variates depends upon the nature and characteristics of the data and the purpose for which it is to be used. In the case of discrete variates, the unit is a natural interval and sometimes it is satisfactory. (See Tables 10 and 11.) However, for both discrete and continuous variates the following conditions should guide the choice: (a) We desire to be able to treat all the values assigned to any one class, without serious error, as if

TABLE 3 -- MONTHLY RAINFALL AT IOWA CITY, 1890-1925

<i>Year</i>	<i>Jan.</i>	<i>Feb.</i>	<i>Mar.</i>	<i>Apr.</i>	<i>May</i>	<i>June</i>	<i>July</i>	<i>Aug.</i>	<i>Sept.</i>	<i>Oct.*</i>	<i>Nov.</i>	<i>Dec.</i>
1890	2.75	0.75	1.80	1.83	2.20	7.99	0.30	2.29	1.44	2.11	1.56	0.31
1891	1.49	1.30	4.41	1.11	4.46	2.80	3.01	3.45	2.33	1.63	2.93	2.72
1892	1.46	1.23	3.15	4.30	9.23	8.29	6.20	2.50	1.18	1.02	1.38	2.84
1893	1.18	1.75	2.82	4.37	1.79	3.01	3.56	1.64	3.07	1.98	1.75	1.52
1894	1.95	1.64	2.03	2.72	3.09	2.40	0.90	2.40	4.96	2.30	1.80	0.98
1895	2.37	0.64	1.25	1.66	4.26	1.10	10.10	1.77	3.43	1.38	1.78	2.84
1896	0.70	1.51	0.92	5.14	4.10	1.86	7.04	2.44	1.82	2.74	1.16	0.55
1897	3.66	1.30	2.07	4.60	3.11	2.38	3.83	1.85	3.54	0.33	1.98	2.48
1898	4.62	1.15	3.02	2.89	4.80	3.26	2.27	2.85	2.54	4.38	1.10	0.53
1899	0.59	1.82	1.43	3.23	9.49	4.50	3.78	2.39	0.93	1.66	1.15	1.93
1900	0.73	2.20	3.32	3.31	4.31	2.18	5.25	6.27	4.35	3.61	1.43	0.75
1901	1.07	1.97	3.62	2.36	1.54	3.33	1.29	0.66	2.56	1.78	0.79	2.34
1902	1.29	0.85	1.29	1.91	3.75	7.46	6.89	10.91	5.87	3.12	2.25	2.21
1903	0.67	1.03	1.86	3.11	6.90	1.95	4.76	3.45	5.38	3.60	0.97	1.27
1904	1.74	0.84	2.73	5.49	2.68	2.14	2.49	3.93	3.12	1.59	0.25	1.96
1905	1.22	1.90	2.28	3.36	5.37	6.68	3.59	2.62	1.54	5.36	2.92	1.04
1906	2.51	1.73	2.25	1.83	2.33	3.64	1.42	5.34	0.89	1.48	3.08	1.64
1907	2.12	0.22	1.59	1.58	5.47	6.04	9.21	2.98	2.85	0.86	1.07	0.53
1908	0.32	2.08	2.94	2.78	7.78	2.87	5.40	7.47	1.82	1.99	1.84	0.43
1909	1.97	1.09	2.00	7.21	4.40	4.58	5.75	1.88	2.43	1.59	4.88	2.52
1910	1.79	0.39	0.28	2.56	3.57	0.98	2.22	4.98	3.87	0.57	0.69	0.46
1911	0.87	4.82	1.30	3.02	4.74	2.98	3.70	4.27	5.07	2.78	3.01	2.29
1912	0.26	1.21	2.30	3.50	2.88	2.60	3.60	3.62	2.67	3.54	1.11	0.75
1913	1.19	1.42	2.69	1.83	6.91	6.28	0.39	2.97	3.19	3.66	0.46	1.02
1914	1.28	0.93	2.63	2.37	4.87	5.32	1.53	2.99	7.97	1.65	0.37	1.89
1915	2.15	2.42	0.92	0.65	7.65	4.33	8.11	1.80	9.31	1.84	1.80	0.80
1916	3.18	0.59	5.06	1.83	5.99	3.92	1.57	2.83	3.49	3.19	1.42	1.15
1917	1.09	0.19	2.19	3.43	7.33	6.49	2.84	2.79	6.23	2.28	0.30	0.57
1918	1.10	1.46	0.33	3.43	6.22	8.36	4.87	6.72	2.00	2.05	2.10	1.62
1919	0.08	2.63	2.65	4.28	4.49	7.07	1.03	2.67	5.10	4.01	3.84	0.61
1920	0.84	1.33	4.22	4.75	3.76	2.86	2.79	2.90	1.20	0.98	1.80	2.45
1921	0.35	0.49	2.46	6.20	4.44	2.46	3.59	8.61	7.83	2.47	0.74	3.19
1922	1.11	1.46	2.18	3.49	5.52	0.28	6.46	1.03	2.91	1.06	5.28	0.49
1923	1.09	0.67	4.83	0.86	2.63	6.21	2.37	4.01	9.27	2.35	1.13	0.73
1924	1.35	0.83	2.10	1.09	1.69	8.71	3.67	5.67	2.60	1.64	0.93	1.75
1925	0.29	1.04	0.99	3.07	1.06	5.61	3.63	3.14	5.59	3.90	1.00	1.66

they were equal to the class mark for that interval; e.g., as if all 23 items in the first class of Table 4 were exactly 0.245 inches, etc.

(b) For convenience and brevity we desire to make the interval as large as possible subject to the first condition. These conditions will generally be fulfilled if the interval is so chosen that the whole num-

TABLE 4 — FREQUENCY TABLE OF MONTHLY RAINFALL AT IOWA CITY,
1890-1925

<i>Class Interval</i>	<i>Mid-x</i>	<i>Frequency</i>
0.00- 0.49	0.245	23
0.50- 0.99	0.745	42
1.00- 1.49	1.245	58
1.50- 1.99	1.745	62
2.00- 2.49	2.245	49
2.50- 2.99	2.745	47
3.00- 3.49	3.245	32
3.50- 3.99	3.745	27
4.00- 4.49	4.245	18
4.50- 4.99	4.745	15
5.00- 5.49	5.245	14
5.50- 5.99	5.745	7
6.00- 6.49	6.245	10
6.50- 6.99	6.745	5
7.00- 7.49	7.245	6
7.50- 7.99	7.745	5
8.00- 8.49	8.245	3
8.50- 8.99	8.745	2
9.00- 9.49	9.245	5
9.50- 9.99	9.745	0
10.00-10.49	10.245	1
10.50-10.99	10.745	1
Total		432

ber of classes lies between 10 and 25. A small number of classes may "cover up" too much detail whereas a large number may reveal too much detail for one to comprehend readily (which is just the objection to the table of original data). A preliminary inspection of the data should accordingly be made and the highest and lowest values selected. Dividing the difference between these by the tentative number of classes, we have our approximate value

TABLE 5 — MONTHLY RAINFALL AT DES MOINES, 1890-1925

<i>Year</i>	<i>Jan.</i>	<i>Feb.</i>	<i>Mar.</i>	<i>Apr.</i>	<i>May</i>	<i>June</i>	<i>July</i>	<i>Aug.</i>	<i>Sept.</i>	<i>Oct.</i>	<i>Nov.</i>	<i>Dec.</i>
1890	2.62	1.17	0.91	0.78	3.00	4.91	1 10	3.35	1.57	4.48	0.74	0.11
1891	1.82	1 13	2 25	2 12	3 29	5 60	2 78	4.22	1 64	2 41	1.34	1.54
1892	1.60	1 35	2.47	3 36	8 77	3 41	8.64	2 45	1.12	2.54	0 76	1.95
1893	0.56	1.28	1 15	5 61	2.84	4 69	3 55	1 60	1.33	0.22	1.51	1.30
1894	1.09	1 39	1 78	1.70	1.41	1 67	0.29	1 89	4.46	2 24	0.99	1.15
1895	1 30	0 60	0 50	3.41	2 86	5.26	3 10	3.57	3.20	0.29	0.85	1.86
1896	0 60	0.79	1.24	3.47	6 50	2 69	8.15	5.49	3.61	2.69	1.10	0.85
1897	2.02	0.71	2.13	7 37	2 31	3 15	2 88	1 77	1 56	0 85	0.34	1.98
1898	1.59	0.82	1.35	2.64	4 22	6 85	1 86	1.09	1.91	3.56	1 87	0.57
1899	0.29	0.57	1.04	2.22	6.71	3 53	3.20	3.53	1 17	0.59	1.76	2.12
1900	0.20	0 50	3 07	3.82	4.76	4 89	5.15	8 02	3.66	3.08	0 96	0.35
1901	1.01	1.11	3.02	2.26	1 40	2.41	1.72	0.67	2 60	2 14	0.40	1.03
1902	0 91	0 52	1.15	1.55	4 69	7 27	5 95	7 82	5 03	3.70	1.65	1.77
1903	0 20	1 12	1 09	1 64	0 64	3 06	3 62	6.72	1.62	1 32	0 31	0.09
1904	1 22	0.22	1 20	5 48	3 16	2 08	6 94	2 60	1.95	1 50	0.06	2 02
1905	1.08	1 00	2.16	3 29	4 44	5 73	4 53	5 21	3 47	3.64	2.34	0.55
1906	2.07	0.86	1 84	2 96	2 21	3 80	2 67	4 69	3 24	1 18	2 29	1.46
1907	0.87	0 93	1.18	1.48	2 97	4 13	10.20	5 03	2 40	1.70	1.12	1.01
1908	0.46	1.15	1.43	2.69	9 89	5 93	1 56	6 54	0.94	3 68	0.95	0.31
1909	1.61	0.90	1.56	5 14	4 24	7.01	4 41	0 14	2 06	2.89	3.71	2.32
1910	1.72	0.20	0.33	1.13	3 26	3.11	0.86	2.40	3.82	0.68	0.53	0.20
1911	0.84	2.91	1.14	4.23	2.44	0.75	1.16	1.82	7.68	2.61	1.22	3 18
1912	0.53	1.86	2.87	2 75	5 62	2 60	3 07	3 52	4.20	3.75	1.11	0.30
1913	1.10	0 65	3.03	3.41	5 06	3 52	1.05	3 44	2.65	2.67	1.03	1.05
1914	0.85	1 24	1.18	1.52	4.83	3 89	1 22	1.77	4.81	3.57	0.35	1.28
1915	1.96	3.20	1.16	1.36	8.21	3.60	9.39	1.71	4.51	0.43	1.24	0.65
1916	2.66	0.61	0.60	2.44	3.87	2 42	1.50	2.62	1.72	2.11	1.46	0.65
1917	6.53	0.52	2.30	5 52	3.94	8.16	1.58	1.82	1.99	0.92	0.21	0.88
1918	0.78	1 45	0.29	1 81	5.87	5.63	1.18	2.54	0.91	3.81	2.10	1.35
1919	0.08	3 00	3.67	5.30	2.96	7.36	2.68	2.19	7.47	2.20	3.84	0.93
1920	0.44	0.74	3.92	4.09	3.14	1.25	5.66	2.11	4.44	1.89	1.63	1.38
1921	0.59	0.92	1.07	3.72	3.62	4.66	2.49	6.63	7.16	1.51	0.35	0.80
1922	0.85	0.64	2.25	2.84	6.87	1.63	7.13	6.63	3.00	3.41	2.54	0.25
1923	0.88	0 36	4.34	1.76	4.78	4.95	0.78	5.34	5.17	1.10	0.55	0.61
1924	1.02	1.98	3.10	0.78	1.26	9.30	0.98	4.15	3.47	0.77	0.53	1.62
1925	0.23	0.50	0.88	1.64	0.77	6.40	2.21	4.79	3.75	3.22	0.32	1.67

of the interval. After a little preliminary reconnoitering an appropriate number of classes and their limits can be determined. Thus, in Table 3, the highest value noted was 10.91 and the lowest 0.08 (verify). The difference between these is 10.83, which suggests that if we took 20 classes we would have approximately a half inch as the width of a class interval. This, however, assumes we would start with 0.08 as our lower limit, which would give us awkward figures as limits. Therefore, our judgment suggests it would be better to start with 0 and continue by half-inch intervals as far as is necessary to take in the range of the given variates. We have estimated it will take approximately 20 of these; actually it turns out to be 22. This number of intervals and their width is consistent with the general conditions (a) and (b) given above. On page 16 are given some supplementary rules which in general are helpful in making a frequency distribution.

8. Distinction between Class Limits and Class Boundaries. The pairs of numbers written in the column of classes of a frequency distribution are the lower and upper *class limits*, sometimes called open class limits. For instance, 1.00–1.49 are the limits of the third class of Table 4. When the measurements of Table 3 were made, readings were recorded to the nearest hundredth of an inch. Thus, a measurement which was more than 1.485 and less than 1.495 was recorded as 1.49. Likewise, if a measurement was more than 0.995 but less than 1.005, it would be recorded as 1.00. Therefore, the third class of Table 4 includes all measurements more than 0.995 and less than 1.495. These values are then the true or closed limits of the third class and are known as *class boundaries* or *end values*. A class boundary is the value halfway between the upper limit of one class and the lower limit of the next class. For example, the upper boundary of the fourth class of Table 4 is 1.995 which is the lower boundary of the fifth class. If we denote the variate values by x , the following table illustrates these remarks for the first five classes of Table 4.

<i>Class Limits</i>	<i>End-x</i>	<i>Mid-x</i>
0.00–0.49	0.495	0.245
0.50–0.99	0.995	0.745
1.00–1.49	1.495	1.245
1.50–1.99	1.995	1.745
2.00–2.49	2.495	2.245

The width of a class interval is the same, however, whether the

classes are expressed in terms of class limits or class boundaries, being the difference between the beginning of one class and the beginning of the next class. Similarly, the class mark as the mid-point of the interval is unaffected. Thus, for the class limits 1.00-1.49, the class mark is $\frac{1}{2}(1.00 + 1.49) = 1.245$; for the corresponding class boundaries, the class mark is $\frac{1}{2}(0.995 + 1.495) = 1.245$.

The distinction between class limits and class boundaries is an important one in plotting graphs, but in tabulating it is the class limits that should be expressed.

9. Rules for Making a Frequency Distribution.

- (1) Determine the range of the table by finding the difference between the highest value and the lowest value among the items.
- (2) Determine the number of equal parts into which the range shall be divided. The size of the class interval and the number of intervals depend upon the size and nature of the distribution. (Table 1 contains rather fewer classes than is usually desirable but an interval of 10 units is quite conventional in students' grades. An interval of 5 would be used if grades of A, A—, B, B—, etc., were given instead of A, B, etc.) Intervals of 0.5, 1, 2, 3, 5, 7, or 10 are the most common.
- (3) Arrange a sheet with three headings: class interval, tally marks, frequency.
- (4) Read off the items in the raw table and for each one record a mark, as shown in Table 2.
- (5) Write the sum of the marks in each row in the frequency column. The sum of the frequencies should, of course, equal the total number of variates.

10. Cumulative Frequencies. The frequencies with which we have been concerned may be called absolute frequencies to distinguish them from two other kinds which will be mentioned in this course; namely, cumulative frequencies and relative frequencies. The first of these will be considered here.

Sometimes a statistical investigation is concerned with the number or percentage of variates which are "less than" or "more than" a given value. This is frequently the case in educational tests and in wage or salary statistics. Our chief interest in such cases may be the accumulated frequency of the several class intervals up to some class boundary. Hence we are led to form a cumulative frequency table. Such a table is built up by successively adding the several

(absolute) frequencies; thus: $f_1, f_1 + f_2, f_1 + f_2 + f_3$, etc., as illustrated in Table 7, where the data of Table 6 are used. We shall use N to denote the sum of all the frequencies.

TABLE 6 — DISTRIBUTION OF INTELLIGENCE QUOTIENTS (IQ's) OF 905 SCHOOL CHILDREN FROM 5 TO 14 YEARS OF AGE. (DERIVED FROM L. M. TERMAN, *The Measurement of Intelligence*)

<i>IQ</i>	<i>Number of Children</i>
55- 64	3
65- 74	21
75- 84	78
85- 94	182
95-104	305
105-114	209
115-124	81
125-134	21
135-144	5

The cumulative frequency (*cum f*) at any class is the total (absolute) frequency up to the upper boundary of that class. This is the reason for placing the *cum f* entries opposite the *end-x* values and on lines between the *mid-x* entries. Thus, in the *cum f* column of Table 7, three students had IQ's less than 64.5, 24 less than 74.5, etc. The

TABLE 7 — CUMULATIVE DISTRIBUTION OF IQ's (TABLE 6)

<i>Class Mark</i> <i>Mid-x</i>	<i>Frequency</i> <i>f</i>	<i>Upper Boundary</i> <i>End-x</i>	<i>Cum f</i>	$\frac{\text{Cum } f}{N}$
59.5	3 = f_1	54.5	0	0.000
69.5	21 = f_2	64.5	3 = f_1	0.003
79.5	78	74.5	24 = $f_1 + f_2$	0.027
89.5	182	84.5	102	0.113
99.5	305	94.5	284	0.314
109.5	209	104.5	589	0.651
119.5	81	114.5	798	0.882
129.5	21	124.5	879	0.971
139.5	5	134.5	900	0.994
		144.5	905 = N	1.000

entries in the column headed (*cum f*)/ N give the percentages of the total frequency which are less than the values of the *end-x* column. Thus, from this column in Table 7, we can readily see that 88% of the children had IQ's less than 114.5 and only 11% less than 84.5.

Table 7 is known as a "less than" table. One could of course cumulate the frequencies from the bottom of the table, getting a "more than" distribution. The *cum f* column would then give the number of children whose IQ's are more than the values at the lower boundaries of the several class intervals.

The inverse operation to cumulating the frequencies is called "differencing" and is usually denoted by Δ (delta). If S denotes any series of values, then ΔS denotes the results obtained by subtracting the first value of S from the second value, the second from the third, etc. Differencing a column of cumulative frequencies obviously gives the absolute frequencies. Differencing a column of $(\text{cum } f)/N$ values gives the f/N values.

Exercises

1. What is the width of the class interval and the values of the class marks in Table 2?
2. Tabulate the grades of Table 1, using class intervals of 5 units.
3. With reference to Table 3, is it easy to answer such questions as the following:
 - (a) In how many instances are the monthly rainfall between 2 inches and 3 inches?
 - (b) In how many instances was the rainfall less than 5 inches?
 - (c) What was the smallest monthly rainfall recorded?
 - (d) What per cent of the total measured between 5 inches and 10 inches?
 - (e) What measurement is the most common?
4. Refer to Table 4 and then answer the above questions.
5. Using your own judgment as to the most appropriate class interval, make a frequency distribution of the monthly rainfall for Des Moines from 1890 to 1925 (Table 5).
6. For Table 6 state the class boundaries (end values) and the class marks.
7. Difference the *cum f* column of Table 7.
8. Read the following references:
 - (a) *Mathematics Essential for Elementary Statistics* — Walker, Chapter II.
 - (b) *Standards and Requirements in Statistics* — Belcher. *Journal American Statistical Association*, vol. 21, p. 424.

11. Additional Distributions. The following distributions which will be referred to in subsequent chapters will serve as illustrative and laboratory material. They are not chosen on account of the importance of the data but merely to exemplify methods.

TABLE 8 — DISTRIBUTION OF LENGTHS OF
995 TELEPHONE CALLS. TIME IN SECONDS

<i>Time</i>	<i>Number of Calls</i>
0-99	1
100-199	28
200-299	88
300-399	180
400-499	247
500-599	260
600-699	133
700-799	42
800-899	11
900-999	5

(For future reference: $\bar{x} = 477.3$ secs., $\sigma = 148.5$ secs.)

TABLE 9 — DISTRIBUTION OF WEIGHT IN POUNDS AMONG
1000 8-YEAR-OLD GLASGOW SCHOOLGIRLS

<i>Weight (mid-values)</i>	<i>Frequency</i>
29.5	1
33.5	14
37.5	56
41.5	172
45.5	245
49.5	263
53.5	156
57.5	67
61.5	23
65.5	3

TABLE 10

Twelve dice were thrown 4096 times; only a throw of 6 was counted a success.
The observed distribution follows:

<i>Successes</i>	<i>Frequencies</i>
x_i	f_i
0	447
1	1145
2	1181
3	796
4	380
5	115
6	24
7	7
8	1
9	0
10	0
11	0
12	0

(For future reference: $\bar{x} = 2$, $\sigma = 1.296$)

TABLE 11

Twelve dice were thrown 4096 times; a throw of 4, 5, or 6 points being reckoned a success. The following distribution was recorded:

<i>Successes</i>	<i>Frequency</i>
0	0
1	7
2	60
3	198
4	430
5	731
6	948
7	847
8	536
9	257
10	71
11	11
12	0

(For future reference: $\bar{x} = 6.139$, $\sigma = 1.712$)

TABLE 12 — FREQUENCY DISTRIBUTION OF THE WEIGHTS OF 1000 MALE STUDENTS (ORIGINAL MEASUREMENTS MADE TO NEAREST HALF POUND)

<i>Class Pounds</i>	<i>Class Mark</i>	<i>Frequency</i>	<i>Cumulative Frequency</i>
90-99.5	94.75	2	2
100-109.5	104.75	21	23
110-119.5	114.75	104	127
120-129.5	124.75	196	323
130-139.5	134.75	248	571
140-149.5	144.75	197	768
150-159.5	154.75	133	901
160-169.5	164.75	47	948
170-179.5	174.75	25	973
180-189.5	184.75	14	987
190-199.5	194.75	7	994
200-209.5	204.75	4	998
210-219.5	214.75	0	998
220-229.5	224.75	0	998
230-239.5	234.75	1	999
240-249.5	244.75	1	1000

(For future reference: $\bar{x} = 138.65$, $\sigma = 18.03$, $\alpha_1 = .94$)

TABLE 13 — DISTRIBUTION OF SPAN (CENTRAL VALUES) IN INCHES AMONG 2000 ADULT MALES (ORIGINAL MEASUREMENTS TO THE NEAREST INCH)

<i>Span</i>	<i>Frequency</i>	<i>Span</i>	<i>Frequency</i>
58.5	1	71.5	217
59.5	2	72.5	176
60.5	1	73.5	132
61.5	6	74.5	82
62.5	7	75.5	48
63.5	22	76.5	20
64.5	55	77.5	16
65.5	111	78.5	12
66.5	146	79.5	3
67.5	182	80.5	1
68.5	229	81.5	2
69.5	265	82.5	1
70.5	263	Total	2000

The following references are recommended to those who desire some distributions which may be more interesting in themselves:

- (a) Per cent Distribution of Deaths in Each Age Period, by Specified Causes. White Males and White Females, United States, 1942. Source: Metropolitan Life Insurance Company, *Statistical Bulletin*, October 1945, p. 7.
- (b) Age of American Military Leaders. Source: Metropolitan Life Insurance Company, *Statistical Bulletin*, June, July, August, 1945.
- (c) Employment Status of the Population by Age and Sex. Source: *Population, Third Series, The Labor Force*, Table 5, 16th Census.
- (d) Distribution of Population by Age. Source: *Statistical Abstract*, 1943, p. 24.

CHAPTER II

GRAPHICAL REPRESENTATION

1. The Function Concept. Variables which are linked or related in some way are encountered in various fields of human experience. Several variables may be linked but we shall, for the present, consider the simple case where only two variables are involved. For example, the two related variables may be time and population, variate and frequency, rate of interest and accumulated principal, age and insurance premium. The primary purpose of a graph is to show diagrammatically how the values of one of two linked variables change with those of the other. One of the most useful applications of the graph occurs in connection with the representation of statistical data.

Underlying the intelligent use of graphs is the concept of *function*, which is a fundamental notion in mathematics and its applications. The mathematical meaning of function is a technical one, entirely different from the ordinary meaning. The student usually meets the word for the first time in algebra, when a linear or quadratic expression is spoken of as a function of x . An example is the equation

$$y = P(1 + x)^2.$$

The expression on the right is the function of x (P being constant) and for convenience it is denoted by the single letter y . Here x is an interest rate and y denotes the amount to which P dollars will accumulate in two years at $x\%$ per year.

The statement that y is a function of x is written symbolically in the form

$$y = f(x).$$

This implies that a value of the function y is determined when a value is assigned to the variable x . For this reason, x is called the *independent variable* and y the *dependent variable*. In place of f other letters may be used. Thus, any one of the symbols

$$g(x), \quad h(x), \quad F(x), \quad \phi(x),$$

and so on, denotes a function of x . The same symbol may be used

to denote different functions in different problems, but different symbols are required to represent different functions in the same problem or discussion.

Examples:

$$f(x) = 5x^2 - 3x + 2,$$

$$\phi(x) = Ke^{-x^2}.$$

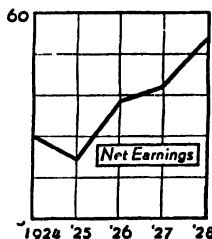
Any mathematical expression involving a variable x is a function of x . However, the word is often used to designate a relation that is completely divorced from any equation or expression. The central idea conveyed by this more general meaning is that of a correspondence between values of y and values of x . The following definition is the result of a development over a long period and its formulation is due to Dirichlet, a famous French mathematician (1805-59).

DEFINITION. *Let there be a set of values assumed by the independent variable x . If to each x in the set, there corresponds one or more values of y , then y is said to be a function of x in the set.*

It should be observed that this definition¹ is freed from any notion of the necessity of specifying the mathematical relation between x and y . We may or may not know the special method by which the correspondence is set up. A mathematical formula or equation between x and y may not even exist. A function may thus be considered as being equivalent to a table in which one may look up any x of the set of the definition, and find the corresponding y .

Much of the data in statistics comes under this general definition of function. Thus, in the following table, net earning is a function of the year, whether or not there is any equation defining that functional relationship.

Here the function is defined only for the indicated points which correspond to the values given in the table. The straight lines are drawn to help the reader visualize the relative positions of these values and not to represent the function at intermediate points. They may, however, be thought of as a first



Year	Millions
1924	45
1925	43
1926	49.6
1927	51.5
1928	57.3

¹ A classical example is the function which is defined for the infinite set of numbers from $x = 0$ to $x = 1$ to be unity for all rational numbers and zero for all irrational numbers.

approximation to the unknown function between the given values. Such a representation of the function could not, of course, be assumed in the case of discrete variates because then the function is discontinuous and does not exist except for the given values.

Referring again to the above definition, if there is only one value of y corresponding to each value of x then y is called a *single-valued* function of x ; otherwise y is said to be a *multiple-valued* function of x . Child weight would be an example of a multiple-valued function of age, being different for different children. The weight of a particular child would be a single-valued function of age. For the most part we shall be concerned with single-valued functions.

2. Charts. A detailed study of the technique of representing data by broken lines, by charts or bar graphs, etc., will not be undertaken here. It is a rather specialized and non-mathematical subject, and the student interested in plain-scale cartography can readily find books on the subject which are very readable.¹ (A discussion of ratio charts is given in Chapter VII.)

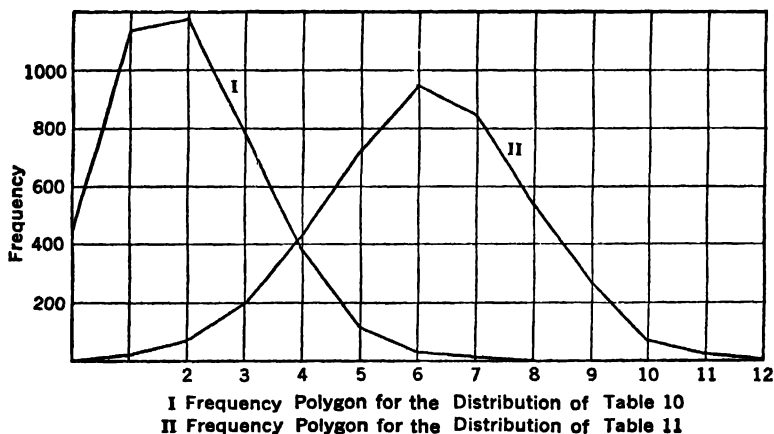


FIG. 1 — FREQUENCY POLYGONS FOR DISTRIBUTIONS OF DISCRETE VARIATES

3. Frequency Polygon. We present now a discussion of the graphs that are used in connection with frequency distributions. A

¹ For example,

- (a) *Graphs: How to Make and Use Them* — H. Arkin and R. Colton. 2nd ed. Harper.
- (b) *Engineering and Scientific Graphs for Publication*. American Standards Association, New York.
- (c) Reference 8 in our Introduction.

distribution of discrete variates may be represented graphically by plotting the points (x_1, f_1) , (x_2, f_2) , \dots (x_k, f_k) , and drawing a broken line through them. Such a graph is called a frequency polygon because it is a polygon formed by connecting the tops of a series of ordinates whose lengths are proportional to the various frequencies and whose abscissas correspond to the variate values of the distribution. Figure 1 will serve as an illustration. For a table of discrete variates the function exists only for the given values. Likewise, its graph is discontinuous. The straight lines connecting the points serve merely to "carry the eye," thus giving a better idea of the shape and position of the distribution.

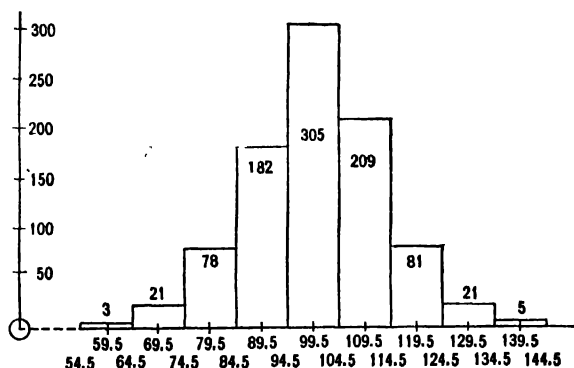


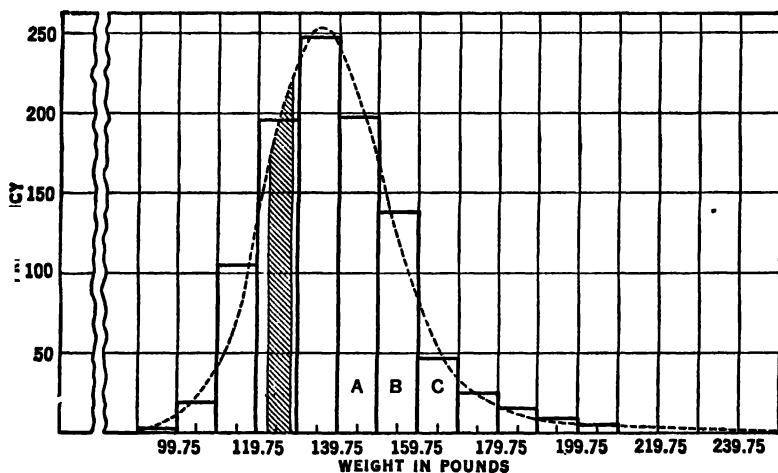
FIG. 2 — HISTOGRAM FOR TABLE 6

4. Histogram. If the frequency distribution is one of grouped variates (discrete or continuous) it is better to use some form of graphical representation which recognizes the fact that the several measurements in a table do not lie precisely at the class marks but are spread out over the intervals of which the class marks are centers. This may be accomplished through the use of a *histogram*. A histogram is a series of rectangles erected at the class boundaries with altitudes proportional to the respective class frequencies, and centered on the class marks. Thus the frequencies are represented by areas. (See Figure 2.) If the bases are all of unit length then the altitudes are also equal to the frequencies. The histogram is an important and useful graphical device for representing frequency distributions.

5. Frequency Curves. The shape of the distribution may be emphasized by constructing a continuous frequency curve such that

the areas under the curve between the ordinates at the upper and lower boundaries of the various rectangles will equal approximately the areas of the corresponding rectangles. Thus, in Figure 3, the area of all the rectangles represents the total frequency 1000, and the area of the three rectangles labeled A, B, C represents the number of individuals weighing between 139.75 pounds and 169.75 pounds. The dotted line represents roughly the frequency curve corresponding to the histogram.

Representing each class frequency of a distribution of continuous variates by a rectangle is equivalent to saying that we realize that



Frequency Distribution of the Weights of 1000 Male Students (Table 12)

FIG. 3 — HISTOGRAM AND FREQUENCY CURVE FOR A DISTRIBUTION OF CONTINUOUS VARIATES

the function exists for points other than the class marks, but we do not know what it is for these points, and so as a first approximation we assume that the variates are uniformly distributed over each interval, which is equivalent to regarding them as concentrated at the class marks. If the class intervals were made smaller and smaller and at the same time the number of variates were proportionally increased, the upper bases of the rectangles would approach more and more the frequency curve which represents the ideal or theoretical mathematical function relating frequency with variate value for the given distribution.

A frequency curve is often drawn for convenience in describing

the properties of an observed distribution, although strictly speaking, the concept of a frequency curve is applicable only to an infinite "universe" of continuous variates. The data at hand are supposed to be a "sample" from the universe represented by the frequency curve.

The more common types of distributions may be represented by bell-shaped curves which are either symmetrical or skew. For elementary purposes it is sufficient to consider frequency distributions as of these two general types. In passing, we may also mention two other types which are known as J-shaped and U-shaped. For examples of these types see Yule and Kendall, *An Introduction to the Theory of Statistics*, Ch. VI.

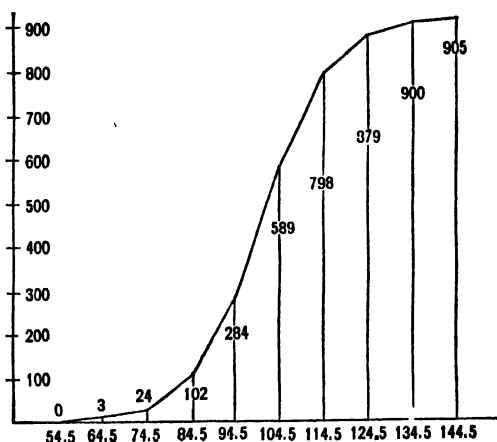


FIG. 4 — OGIVE FOR TABLE 7

6. Ogives. The graphs of cumulative frequencies are called *ogives*. The ogive for Table 7 is shown in Figure 4 and is constructed by plotting the points (54.5, 0), (64.5, 3), etc., as in algebra, and joining them with straight lines.

The student should observe that while $\text{cum } f$ is a function of x it is defined for the *end- x* values only. Occasions will perhaps arise when we desire the x -value corresponding to some intermediate $\text{cum } f$ value, say 453 in Figure 4. Conversely, we might wish to know the $\text{cum } f$ value for some intermediate x -value, say at $x = 97$. Strictly speaking, we do not know the answer in either case, inasmuch as we do not know how the IQ's are distributed over the interval. Perhaps all the individual values in the interval 94.5–104.5 (say) are

less than 97; perhaps none are. The fairest assumption we can make is that they are uniformly distributed throughout the interval. This means graphically that we represent *cum f* over each interval by a straight line, as is done in Figure 4. We may now interpolate under this line for intermediate values. This is "straight line interpolation" and is what the student uses when he interpolates in logarithms.

More refined methods exist for interpolating values of a function between the observed values but their study constitutes a separate branch of mathematics beyond the scope of this course. It should be observed that the straight line used here is a first approximation to the unknown function, and not merely a device to carry the eye as in the case of a frequency polygon for a discontinuous distribution of discrete variates.

7. Relation of *Cum f* to Areas. The sum of the frequencies (*cum f*) up to any value of *x* means, graphically, the sum of the areas of the rectangles of the histogram up to that value. Thus in Figure 4, the ordinate erected at $x = 84.5$ represents the sum of the frequencies $(3 + 21 + 78) = 102$ (Figure 2). If a frequency curve represents the distribution, then *cum f*, corresponding to any value of *x*, is the area under the curve up to that value. Thus, in Figure 3, *cum f* corresponding to $x = 139.75$ is approximately the area under the smooth curve up to $x = 139.75$, and the total area under the curve is *cum f* = *N*.

Exercises

- If $f(x) = 2x^3$ exhibit $f(-x)$. Give the value of $f(3)$, of $f(-2)$.
 - Let $f(x)$ denote a given function which is defined for all real values of *x* under consideration so that if *c* is any admissible number $f(c)$ is defined. What is the graphical meaning of $f(c)$?
- If $\phi(x) = Ke^{-x}$, (a) show that $\phi(x) = \phi(-x)$; (b) give the value of $\phi(0)$.
- If $h(x) = ax^2 + bx + c$, and $h(x) = h(-x)$, show that $b = 0$.
- If $f(x) = x^2$, show that $f(u) \times f(v) = f(u + v)$.
- If $g(x) = \log\{(1 - x)/(1 + x)\}$, show that $g(u) + g(v) = g\{(u + v)/(1 + uv)\}$.
- Make a histogram for the data of Table 4.
- Same as exercise 6 for Table 8 or 9.
- Construct an ogive for the cumulative frequencies given in Table 12.
- Find the cumulative frequencies and construct the ogive for Table 9.
- For further discussion of ogive curves and their uses, read the following references:

- Elements of Statistics* — Davis and Nelson, pp. 23–28.
- The Mathematics of Statistics* — Burgess, pp. 61–72.

CHAPTER III

AVERAGES

1. It was pointed out in Chapter I that classification of the variates of any long series is the first step necessary to overcome the confusion of detail in the original observations, and to make comparisons with other distributions possible. In Chapter II graphical methods were studied which describe, to some extent, the shape and position of the distribution. Although these methods are helpful, their contribution is largely qualitative.

It is desirable to formulate quantitative descriptions for characterizing a distribution, and as an aid in this direction *averages* are very useful. They are also called *measures of location*. An average is a quantity locating a central value of the distribution. In a sense, it is a typical value of the whole set of variates, although it is not necessary that it actually have the value of one of the items of the set it represents. There are five averages in common use. These are: *arithmetic mean*, *mode*, *median*, *geometric mean*, and *harmonic mean*. The means and median are most frequently used although the arithmetic mean is by far the most important in general statistical work, and the others are of service in special cases. We will consider them in the order named. First, however, it will be desirable to discuss certain symbols and notation which will facilitate the development of formulas.

2. **Notation.** If x denotes a variable, then x_1, x_2, \dots, x_N , are general symbols for the values which x may take. When we are concerned with a sum like the following,

$$x_1 + x_2 + x_3 + x_4 + \dots + x_i + \dots + x_N,$$

it is customary to designate it by placing the Greek capital letter Σ (sigma) before the general term, thus

$$\sum_{i=1}^N x_i = x_1 + x_2 + \dots + x_i + \dots + x_N.$$

The symbol Σ is a sort of mathematical verb and the notation written above and below it may be called adverbs. Mathematicians

call \sum an *operator* and speak of the "adverbs" as *limits*. When $\sum_{i=1}^N$ is placed before any quantity, it means, "add up all quantities like \dots which are formed by giving i the values of every positive integer from $i = 1$ to $i = N$, inclusive." Thus if x_i stands for "variates" in Table 1, x_1 refers to the first value 75, x_2 refers to the second value 80, etc., and x_N refers to the last value 56. Here $N = 100$. Hence the compact notation $\sum_{i=1}^{100} x_i$ denotes the sum of all the variates in

Table 1. The symbol $\sum_{i=1}^N x_i$ is read, "the summation of x -sub- i , i varying (or running) from one to N ." The subscript i is called the *index* of summation. Any letter may be used as an index but it is conventional to use i or j . Also the upper limit may be denoted by any letter but we shall use N to denote the total number of variates (some of which may be alike) in a set.

If a variable x is to take on the particular values, 1, 2, 3, etc., instead of the general values x_1, x_2, x_3 , etc., then x itself becomes the index of summation and we write $x = 1$ underneath \sum . Thus

$$\sum_{x=1}^N x = 1 + 2 + 3 + \dots + N,$$

$$\sum_{x=1}^N x^2 = 1 + 2^2 + 3^2 + \dots + N^2.$$

Frequently the index of summation is understood from the context and the notation at the top and bottom of \sum may be omitted if no ambiguity results.

It is imperative that the student master, as soon as possible, the significance and utility of the \sum notation.

Illustrations:

1. $\sum_{i=1}^N 3x_i = 3x_1 + 3x_2 + \dots + 3x_N$
 $= 3(x_1 + x_2 + \dots + x_N).$
2. $\sum_{i=1}^5 (x_i + c) = (x_1 + c) + (x_2 + c) + (x_3 + c)$
 $+ (x_4 + c) + (x_5 + c)$
 $= (x_1 + x_2 + x_3 + x_4 + x_5) + 5c.$
3. $\sum_{i=1}^4 x_i f_i = x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4.$

$$4. \sum_{j=1}^n x_j y_j = x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4.$$

$$5. \sum_{u=1}^N u^2 = 1^2 + 2^2 + 3^2 + \cdots + N^2.$$

The following simple theorems will be useful in our work.

Theorem I. *The summation \sum of an algebraic sum of two or more terms is the same algebraic sum of the \sum 's of these terms taken separately. In symbols:*

$$\sum_{i=1}^N (x_i + y_i - z_i) = \sum_{i=1}^N x_i + \sum_{i=1}^N y_i - \sum_{i=1}^N z_i.$$

Theorem II. *A constant factor may be removed from under the summation sign and written outside as a factor. Thus,*

$$\sum_{i=1}^N c x_i = c \sum_{i=1}^N x_i.$$

Proofs: It is left as an exercise for the student to prove these two theorems by expanding the expressions.

Theorem III. *If the expression under $\sum_{i=1}^N$ is a constant c , the expanded result is Nc .*

Examples:

$$1. \sum_{i=1}^N c = c + c + \cdots + c = Nc.$$

$$\begin{aligned} 2. \sum_{i=1}^N (x_i - c) &= \sum_{i=1}^N x_i - \sum_{i=1}^N c, \text{ by Theorem I} \\ &= \sum_{i=1}^N x_i - Nc, \text{ by Theorem III.} \end{aligned}$$

The above theorems hold also if we replace the notation

$$\sum_{i=1}^N x_i \text{ by } \sum_{x=1}^N x, \text{ etc.}$$

The next two theorems have to do with summing integers. The numbers used in counting,

$$1, 2, 3, 4, 5, \dots$$

are called integers or natural numbers.

Theorem IV. *The sum of the first N integers is*

$$\frac{N(N+1)}{2}.$$

In symbols:
$$\sum_{x=1}^N x = \frac{N(N+1)}{2}.$$

This result follows from the fact that the integers form an arithmetic progression.

Theorem V. *The sum of the squares of the first N integers is*

$$\frac{N(N+1)(2N+1)}{6}.$$

In symbols:
$$\sum_{x=1}^N x^2 = \frac{N(N+1)(2N+1)}{6}.$$

Proof: Let us take the identity $x^3 - (x-1)^3 = 3x^2 - 3x + 1$, and sum each side for $x = 1$ to N . Thus,

$$\sum_{x=1}^N [x^3 - (x-1)^3] = \sum_{x=1}^N [3x^2 - 3x + 1].$$

Applying Theorems I-III to the right member we have

$$\sum_{x=1}^N [x^3 - (x-1)^3] = 3 \sum_{x=1}^N x^2 - 3 \sum_{x=1}^N x + N.$$

Performing the indicated sum in the left member, we have

$$\left. \begin{array}{l} 1^3 - 0^3 \\ 2^3 - 1^3 \\ 3^3 - 2^3 \\ \vdots \\ N^3 - (N-1)^3 \end{array} \right\} \text{whose sum is } N^3.$$

Therefore
$$N^3 = 3 \sum x^2 - 3 \sum x + N.$$

Hence, using Theorem IV and simplifying,

$$\sum_{x=1}^N x^2 = \frac{2N^3 + 3N(N+1) - 2N}{6},$$

whence

$$\sum_{x=1}^N x^2 = \frac{N(N+1)(2N+1)}{6}$$

3. Arithmetic Mean. The arithmetic mean of a set of variates is defined as the sum of the variates divided by their number. We are thinking now of a set of ungrouped variates, like that of Table 1. If we use the symbol \bar{x} to represent the arithmetic mean of the N variates $x_1, x_2, x_3 \dots, x_N$, then

$$\bar{x} = \frac{1}{N} (x_1 + x_2 + x_3 + \dots + x_N),$$

or using the more compact notation of the preceding section, we have

$$(1) \quad \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i.$$

Each item in the set is thus represented in the arithmetic mean in proportion to its magnitude.

As an illustration, it is easily verified that for the set of grades given in Table 1,

$$\bar{x} = \frac{7267}{100} = 72.67.$$

Computing the mean¹ strictly according to definition (1) may be called the serial method to distinguish it from other methods which will be presented. This definition is applicable when N is so small that a grouping of the variates into a frequency distribution is not feasible.

If x refers to the integers from 1 to N their mean is

$$(1a) \quad \bar{x} = \frac{1}{N} \sum_{x=1}^N x.$$

4. Weighted Arithmetic Mean. It will be noticed that several of the grades given in Table 1 are alike. For example, 80 occurs seven times. It should be evident that the same result would be found for the mean if, instead of summing the individual values, each value was first multiplied by the frequency with which it occurs and all such products were then added. In general, if the values x_1, x_2, \dots, x_k occur with corresponding frequencies f_1, f_2, \dots, f_k , respectively,

¹ When there is no ambiguity, the arithmetic mean is often referred to as the mean.

where $f_1 + f_2 + \cdots + f_k = N$, it follows that

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{f_1 + f_2 + \cdots + f_k},$$

or, in shorter notation

$$(2) \quad \bar{x} = \frac{1}{N} \sum_1^k f_i x_i, \text{ where } N = \sum_1^k f_i.$$

When obtained in this way, \bar{x} is generally called a weighted arithmetic mean. The term originated in experimental science where some readings which have been made under more favorable conditions are "weighted" according to their reliability or importance. When the weights have been chosen, they become, essentially, frequencies.

If the x 's are added individually, the f 's become unity, and equation (2) reduces to (1). The student should notice that, for the same data, $\sum_1^k f_i x_i$ is numerically equal to $\sum_1^N x_i$. He should also observe that N refers to the number of variates in the set (some of which may be alike), whereas k refers to the number of *different* values of x in the set and hence to the number of products of the form $x_i f_i$ where f_i is the number of times x_i occurs. In the following example, $N = 8$ and $k = 4$.

Example. For the values 6, 8, 7, 6, 5, 7, 6, 5,

$$\sum_{i=1}^8 x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 = 6 + 8 + 7 + 6 + 5 + 7 + 6 + 5 = 50.$$

$$\sum_{i=1}^4 f_i x_i = f_1 x_1 + f_2 x_2 + f_3 x_3 + f_4 x_4 = 2 \cdot 5 + 3 \cdot 6 + 2 \cdot 7 + 1 \cdot 8 = 50. \quad \sum_{i=1}^4 f_i = 8.$$

By either method, $\bar{x} = 50/8 = 6.25$.

Exercises

1. Write in expanded form:

$$(a) \sum_{i=1}^k x_i f_i; \quad (b) \sum_{i=1}^k x_i^2 f_i; \quad (c) \sum_{i=1}^k (x_i - \bar{x}) f_i.$$

2. Write in expanded form:

$$(a) \sum_1^{n_1} f_i; \quad (b) \sum_{i=n_1+1}^{n_1+n_2} f_i; \quad (c) \sum_{i=1}^{n_1} x_i f_i + \sum_{i=n_1+1}^{n_1+n_2} x_i f_i$$

3. Express 2(c) as a single summation, if $n_1 + n_2 = k$.

4. Write in the abbreviated form, using \sum :

(a) $x_1f_1 + x_2f_2 + \cdots + x_kf_k$.

(b) $(x_1 - \bar{x})f_1 + (x_2 - \bar{x})f_2 + \cdots + (x_k - \bar{x})f_k$.

(c) $\frac{1}{N} [(x_1 - \bar{x})^2f_1 + (x_2 - \bar{x})^2f_2 + \cdots + (x_k - \bar{x})^2f_k]$.

5. Prove:

(a) $\sum_{i=1}^k (x_i + 1)^2f_i = \sum_{i=1}^k x_i^2f_i + 2\sum_{i=1}^k x_if_i + N$.

(b) $\sum_{x=0}^n x(x-1)p = \sum_{x=2}^n x(x-1)p$.

6. Compute the value of exercise 1(c) for the example in §4, using the following form:

x_i	f_i	$(x_i - \bar{x})$	$(x_i - \bar{x})f_i$
5	2	-1.25	-2.50
6	3		
7	2	?	?
8	1		

$$\sum (x_i - \bar{x})f_i = ?$$

7. Distinguish between $\sum_{i=1}^N x_i y_i$ and $\left(\sum_{i=1}^N x_i\right)\left(\sum_{i=1}^N y_i\right)$. Write in expanded form.

8. (a) Express in \sum notation: Each different variate is multiplied by its own f and the sum of the results is divided by N .

(b) Give word statements of the expressions in Exercise 4.

(c) Express the general polynomial of degree n in x ,

$$a_0 + a_1x + a_2x^2 + \cdots + a_n$$

in \sum notation.

9. Using the identity

$$x^2 - (x-1)^2 = 2x - 1$$

derive the result

$$\sum_{x=1}^N x = \frac{N(N+1)}{2}$$

by a method analogous to the proof of Theorem V.

10. (a) Express in abbreviated notation: The sum of the squares of the x 's divided by the square of their sum.

(b) If x refers to the integers from 1 to N , evaluate your answer to (a) in terms of N .

(c) Show that the mean of the first N integers is $(N+1)/2$.

5. Arithmetic Mean from Frequency Table. The variates in each class interval of a frequency distribution are assumed to have the value of the class mark for that interval. Therefore, we may use

formula (2) to find the mean of a frequency distribution. In this case, x_i represents the mid-value of the i th class interval, f_i the corresponding frequency, and k the number of intervals; i running from 1 to k . The method of applying (2) is illustrated in Table 14 from the data of Table 2.

TABLE 14

<i>Class Interval</i>	<i>Class Mark x</i>	<i>Frequency f</i>	<i>Product fx</i>
30-39	34.5	2	69.0
40-49	44.5	3	133.5
50-59	54.5	11	599.5
60-69	64.5	20	1290.0
70-79	74.5	32	2384.0
80-89	84.5	25	2112.5
90-99	94.5	7	661.5
Totals		$\Sigma f = 100$	$\Sigma fx = 7250.0$

$$\bar{x} = \frac{7250}{100} = 72.50.$$

If we denote the class interval by c then it is obvious that $c = 10$ in Table 14.

In this connection it is interesting to note that our result here differs very little from the true value 72.67 and therefore our assumption that all values in a given class may be taken as the class mark seems to cause little error in the result obtained for the mean. This can be proved mathematically (under certain assumptions) and will be referred to later.

6. Translation of Axes; Deviations. It is frequently useful to employ the methods and results of geometry in connection with the problems of statistics. Foremost among these methods is the representation of numbers by points on a line; an origin and a unit of measure having been chosen, a coördinate is assigned to each point on the line. When a frequency distribution is represented by a graph, we have seen in Chapter II that the variate values are used as abscissas or measurements along the x -axis. The mean is therefore the point on the x -axis whose coördinates are $(\bar{x}, 0)$. Its position may be emphasized by drawing a vertical line through this point, but it is

the horizontal distance of the point from the origin and not the vertical line which represents graphically the mean.

In discussing the variates we may often work with smaller numbers by changing the origin of reference. If new axes, $x'y'$, are taken parallel to the old axes, xy , with positive directions preserved, the axes are said to be *translated* from one position to the other. A translation of axes corresponds to a transformation of coördinates. Thus if we let

$$x' = x - x_0, \quad y' = y - y_0$$

the origin is translated to the point (x_0, y_0) . Since the variates are denoted by x we are concerned here only with the transformation $x' = x - x_0$ which translates the origin to the point $(x_0, 0)$. The variates referred to a new origin are often called *deviations*. In particular if we translate the origin to the mean by letting

$$x' = x - \bar{x},$$

then for a frequency distribution the deviations are the values obtained by subtracting \bar{x} from each of the class marks. Thus,

$$x_1' = x_1 - \bar{x}$$

$$x_2' = x_2 - \bar{x}$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$x_k' = x_k - \bar{x}.$$

The units of measurement remain unchanged. Figure 5 shows the two systems when the axes are translated to $(\bar{x}, 0)$. Obviously, any

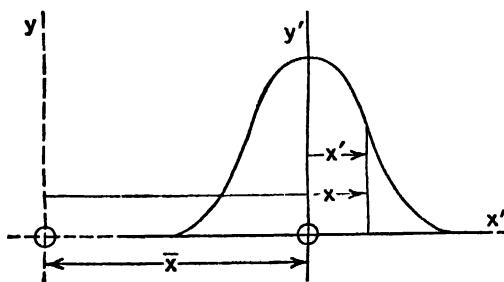


FIG. 5

variates that are larger than \bar{x} will be positive in terms of x' and any variates smaller than \bar{x} will be negative in terms of x' .

7. Properties of \bar{x} . There are two important properties of \bar{x} which may be stated in the following theorems:

Theorem VI. *The algebraic¹ sum of the deviations of all the variates from their arithmetic mean is zero.*

Proof: Let x' represent a deviation from the mean. Multiplying each different deviation by the number of times it occurs and adding these products we have,

$$\begin{aligned}\sum_1^k f_i x_i' &= \sum_1^k f_i (x_i - \bar{x}) \\ &= \sum_1^k f_i x_i - \sum_1^k f_i \bar{x}, \text{ by Theorem I} \\ &= \sum_1^k f_i x_i - \bar{x} \sum_1^k f_i, \text{ by Theorem II.}\end{aligned}$$

Recalling from (2) that $\sum_1^k f_i x_i = N\bar{x}$, and that $\sum_1^k f_i = N$, we have

$$(3) \quad \sum_1^k f_i (x_i - \bar{x}) = N\bar{x} - \bar{x}N = 0.$$

Theorem VII. *If the variates are referred to a new origin x_0 and expressed in units of c by means of the transformation*

$$(4) \quad u = \frac{x - x_0}{c}, \quad (c \neq 0),$$

then the old mean, \bar{x} , is related to the new mean, \bar{u} , by the following formula:

$$(5) \quad \bar{x} = c\bar{u} + x_0.$$

Proof: From (4),

$$(4a) \quad x = cu + x_0$$

and substitution of this value for x in definition (2) gives

$$\bar{x} = \frac{1}{N} \sum_1^k f_i (cu_i + x_0).$$

By Theorems I and II this equals

$$\frac{c}{N} \sum_1^k f_i u_i + \frac{x_0}{N} \sum_1^k f_i.$$

¹ That is, taking account of signs. Some of the deviations will be positive and some negative.

But the first of these expressions is, by definition, c times the mean value of u , and the second is, from (2), simply x_0 . Therefore

$$\bar{x} = c\bar{u} + x_0.$$

This is an important relation and its derivation should be mastered. Observe that the size of a u -unit will be c times as large as the size of an x -unit.

COROLLARY. *If the mean of the deviations of the variates from any arbitrary number, x_0 , is found and added algebraically to x_0 , the result is the mean \bar{x} . In symbols,*

$$(6) \quad \bar{x} = \frac{1}{N} \sum_1^k f_i(x_i - x_0) + x_0.$$

The proof follows from (4) and (5).

In (5) and (6), x_0 may be regarded as a provisional mean, and the first term in the right members may be thought of as the correction to be added algebraically to the provisional mean in order to get the true mean.

8. Short Methods of Computing \bar{x} . In certain cases, the method of computing the mean by (2), as shown in Table 14, can be simplified by use of Theorem VII.

Case I (class intervals equal). If the class marks are equispaced, let c equal the class interval and choose x_0 as one of the class marks, usually the one opposite the largest frequency. From (4), x_0 becomes the origin of u , because when $x = x_0$, $u = 0$.

The method of using (5) is illustrated in Table 15, page 40. Here $c = 10$ and we choose $x_0 = 74.5$, so (4) becomes

$$u = \frac{x - 74.5}{10}.$$

Substituting the given values of x in this relation we get the values in the u column. So in running the fu column, small values of u are multipliers of the larger values of f . Then

$$\bar{u} = \frac{1}{100} \sum fu = \frac{-20}{100} = -.2,$$

so from (5),

$$\bar{x} = 10(-.2) + 74.5 = 72.5\%.$$

It should be evident that the final value obtained for the mean is independent of the choice of the arbitrary value x_0 . This choice is only a rough guess and it is really immaterial which of the given values is selected as x_0 , except that the nearer it is to the mean the lighter will be the calculations to follow. A *check* on the arithmetic may, therefore, be effected by selecting a different provisional mean.

TABLE 15 — MEAN OF 100 GRADES USING CLASS INTERVAL AS UNIT

x	u	f	fu
34.5	-4	2	- 8
44.5	-3	3	- 9
54.5	-2	11	-22
64.5	-1	20	-20
74.5	0	32	0
84.5	1	25	25
94.5	2	7	14
Totals		100	-20

This indirect method is sometimes called *coding* because the variates are coded to another scale in which it is easier to compute the mean. Formula (5) is the relation, then, for transforming the mean from one scale to another.

If one's statistical interests are limited to computing means, then (2) cannot be improved upon if calculating machines are to be used. It should be understood, however, that techniques must be developed now for subsequent purposes. The indirect method is part of a pattern which is useful in later chapters. From this standpoint, one should practice using it at this stage when $N = \sum_1^k f_i$ is large and the x 's are equispaced.

Case II (class intervals unequal). Occasionally a frequency distribution is encountered in which the variates are not equispaced; it is then usually best to take $c = 1$ (unless the x 's have a common factor c) and be content with whatever simplification results from a suitable choice of x_0 . This is equivalent to using the above corollary.

In Table 16, we choose $x_0 = 200$ and are thus able to simplify the work a little. (See page 41.)

TABLE 16

x	f	u	uf
106.12	7	- 93.88	- 657.16
191.83	14	- 8.17	- 114.38
246.48	32	46.48	1487.36
283.63	49	83.63	4097.87
257.65	55	57.65	3170.75
294.51	54	94.51	5103.54
222.53	35	22.53	788.55
71.43	14	- 128.57	- 1799.98
Totals	260		12076.55

$$\begin{aligned}\bar{u} &= \frac{1}{N} \sum f_i u_i = \frac{1}{N} \sum f_i (x_i - 200) \\ &= \frac{12076.55}{260} = 46.448 \\ \bar{x} &= \bar{u} + x_0 = 246.45.\end{aligned}$$

9. Geometric Explanation. Let us consider further the relation between the variables x and u , defined by the expression.

$$(4) \quad u = x - x_0$$

A geometric explanation will be helpful.

Graphically, the x values are distances along the x -axis measured from zero as origin. Likewise x_0 is some point on the x -axis at a distance of x_0 units from zero. If now the points representing the x values are measured from x_0 as origin they are denoted by $x - x_0$.

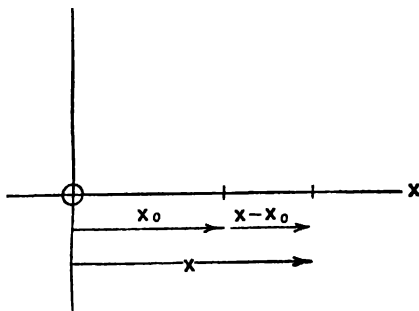


FIG. 6

(See Figure 6.) Thus if $x_0 = 24$, a value which is 36 with reference to the origin of x will be 12 with reference to x_0 ; likewise a value $x = 18$ becomes $x - x_0 = -6$ when referred to x_0 as origin. It

should be noted that $x - x_0$ is in the same units as x . Thus if x is in inches, $x - x_0$ will also be in inches. But $(x - x_0)/12$ would then be in feet. Instead of dividing by 12 suppose we divide by c . Then $(x - x_0)/c$ will be in units of c whatever c may be. It is convenient to denote the resulting values by a different letter, say u . Therefore the numerator of (4) changes the origin of reference but does not affect the scale of measurement. The denominator changes the scale, there being c of the x units in one of the u units. Relation (4) has this generalized meaning apart from statistics. Mathematical notation is applicable to many different fields of knowledge. A relation like (4) which occurs in physics is $C = (5/9)(F - 32)$; it connects temperature on the Centigrade and Fahrenheit scales.

When (4) is applied to a frequency distribution it is convenient to select x_0 as one of the *mid- x* values and to take c as the width of the

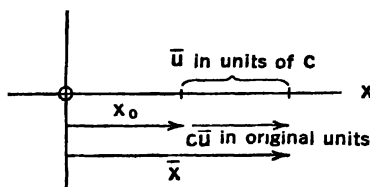


FIG. 7 — If $x_0 < \bar{x}$, $c\bar{u}$ is positive; if $x_0 > \bar{x}$, $c\bar{u}$ is negative.

class intervals. Under Case I, the mean is found with reference to x_0 and in units of c . This is the mean, \bar{u} , of the numbers representing the various class intervals weighted with the corresponding frequencies. After this mean is computed it may be converted back into units of x by multiplying by c , and then referred to the origin of x by adding x_0 . (See Figure 7.) Hence we have $\bar{x} = c\bar{u} + x_0$. Thus we arrive at the same result as that obtained algebraically.

If we had denoted the variates by y we could have used the relation

$$v = \frac{y - y_0}{c}$$

corresponding to (4). Geometrically, this would mean a change of units and a translation of origin in the y -direction. The relation corresponding to (5) would then be

$$\bar{y} = c\bar{v} + y_0$$

where $\bar{v} = \frac{1}{N} \sum f_i v_i$.

As the short-cut method is an important one, another illustration is given in Table 17 (based on Table 4). Here we take $u = (x - 2.745)/0.5 = 2(x - 2.745)$.

TABLE 17 — COMPUTATION OF MEAN MONTHLY RAINFALL AT IOWA CITY
1890-1925

x	f	u	fu
0.245	23	-5	-115
0.745	42	-4	-168
1.245	58	-3	-174
1.745	62	-2	-124
2.245	49	-1	-49
2.745 ← x_0	47	0	0
3.245	32	1	32
3.745	27	2	54
4.245	18	3	54
4.745	15	4	60
5.245	14	5	70
5.745	7	6	42
6.245	10	7	70
6.745	5	8	40
7.245	6	9	54
7.745	5	10	50
8.245	3	11	33
8.745	2	12	24
9.245	5	13	65
9.745	0	14	0
10.245	1	15	15
10.745	1	16	16
Totals	432		49
$\bar{x} = 2.745 + \frac{(0.5)(49)}{432}$ $= 2.802 \text{ inches.}$			

10. Mean of Means. So far we have used subscripts to distinguish between the variates within a set: x_1, x_2, \dots, x_N . By this time the student should be thinking easily in this notation so we may now state an additional use of subscripts. Instead of using x and y to distinguish between two sets of variates we may use x_1 and x_2 . Then to distinguish the variates within a set we would add a second subscript, so for the x_1 set the variates are

$$x_{11}, x_{12}, x_{13}, \dots, x_{1n_1}$$

and for the x_2 set the variates are

$$x_{21}, x_{22}, x_{23}, \dots, x_{2n_2}.$$

These are read "x two one," etc., not "x twenty-one," etc. In the notation dealing with one set, x was a variable but x_1, x_2 , etc., were constants. Now x_1 and x_2 are variables and $x_{11}, x_{12}, \dots, x_{21}, x_{22}, \dots$, etc., are constants. Thus x_1 and x_2 may denote the grades of two sections of mathematics in which there are n_1 and n_2 students respectively. Then the mean of the first set is

$$(a) \quad \bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}$$

and the mean of the second set is

$$(b) \quad \bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2i}.$$

We will now state a useful theorem.

Theorem VIII. *If the mean of a set of n_1 variates is \bar{x}_1 and the mean of another set of n_2 variates is \bar{x}_2 , the mean \bar{x} of the combined sets is*

$$(7) \quad \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{N}$$

where $N = n_1 + n_2$.

Proof: It is obvious from equations (a) and (b) that

$$(c) \quad n_1 \bar{x}_1 + n_2 \bar{x}_2 = \sum_1^{n_1} x_{1i} + \sum_1^{n_2} x_{2i}.$$

If x is allowed to stand for x_1 and x_2 in succession as shown in the table on page 45 then the right member of (c) may be written $\sum_1^{n_1+n_2} x_i$ which denotes the sum of all the variates when they are combined into one set. If this latter sum is divided by the total number of variates N the result is, by definition, their mean. Hence

$$\frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{\sum_1^{n_1} x_{1i} + \sum_1^{n_2} x_{2i}}{n_1 + n_2} = \frac{\sum_1^{n_1+n_2} x_i}{N} = \bar{x}.$$

We may express (7) in more compact notation as follows.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^2 n_i \bar{x}_i, \quad N = \sum_{i=1}^2 n_i.$$

1	x_{11}	x_1	x
2	x_{12}	x_2	
3	x_{13}	x_3	
.	.	.	
.	.	.	
.	.	.	
n_1	x_{1n_1}	x_{n_1}	
$n_1 + 1$	x_{21}	x_{n_1+1}	
$n_1 + 2$	x_{22}	x_{n_1+2}	
$n_1 + 3$	x_{23}	x_{n_1+3}	
.	.	.	
.	.	.	
.	.	.	
.	.	.	
$n_1 + n_2$	x_{2n_2}	$x_{n_1+n_2}$	
$\sum_{i=1}^{n_1} x_{1i} + \sum_{i=1}^{n_2} x_{2i}$		$\sum_{i=1}^{n_1+n_2} x_i$	

This form lends itself to a generalization for k sets so we have the following theorem.

Theorem IX. *The mean of a set of N variates which is composed of k subsets is*

$$(8) \quad \bar{x} = \frac{1}{N} \sum_{i=1}^k n_i \bar{x}_i$$

where \bar{x}_i is the mean and n_i is the frequency in the i th subset and $N = \sum_{i=1}^k n_i$.

COROLLARY. *If $n_i = n$ is the same for all the sets, then $N = kn$ and (8) reduces to*

$$(9) \quad \bar{x} = \frac{1}{k} \sum_{i=1}^k \bar{x}_i$$

Exercises

- (a) Use (1), §3, to find the mean of the following numbers: 18, 42, 23, 16, 103, 61, 49, 95, 113, 10.
(b) For the numbers in (a) verify that the sum of their deviations from their mean is zero. . What theorem does this exercise illustrate?
- Find the deviations of the numbers in Ex. 1 from 50 and verify that the mean of these deviations added algebraically to 50 gives the mean of the numbers themselves.
- Prove: The sum of the deviations of the variates from their mean is zero.
- Derive the relation $\bar{x} = c\bar{u} + x_0$.

5. Find the arithmetic mean of the weights of 1000 students given in Table 12. Use (5). *Ans.* 138.65 lbs.
6. Find the mean monthly rainfall at Des Moines from 1890 to 1925, using the frequency distribution which you previously made. *Ans.* 2.55 inches.
7. Find the mean of the distribution of discrete variates given in Table 11.
8. Prove the following theorem: *The mean of a set of variates is unchanged if each variate is replaced by the mean of all the variates.*
9. (a) Prove expressions (8) and (9).
(b) The mean grade of one class of 20 students is 76% and of another class of 15 students is 80%. Find the mean of the two classes.
10. The record of freshman scholastic averages for a semester at a certain university were given as follows:

	n_i	x_i
Men	501	3.550
Women	356	3.639

Find the mean grade for the entire class.

11. Assume that the following fictitious data represent the earnings per week of a certain type of machine shop labor in Illinois establishments:

Wage Group		Frequency
\$00.0 under \$10.0		50
10.0	20.0	150
20.0	30.0	400
30.0	40.0	200
40.0	50.0	160
60.0	80.0	40
Total		1000

*Class omitted. Note the different interval in the last class.

The average earnings per week for this same type of labor in all other states of the United States where 9000 men are employed, not counting those in Illinois, are \$30.00 per week.

Compute the arithmetic mean wage (a) for Illinois, (b) for the entire United States.

Recompute the mean wage for Illinois in such a manner as to check, in the quickest and surest way, the accuracy of the result found in (a) above.

12. Find the mean of the following distribution:

x	f
47.5	7
48.1	17
45.9	46
44.0	44
40.7	54
41.6	43
38.0	35
33.2	14

11. The Mode. That value of the variable which occurs most frequently is called the *mode*. Its chief service is in characterizing a type and it is the kind of average meant by such a phrase as the "average man." There is some difficulty in giving a precise definition of the mode without more advanced mathematics. However, we may say that for a given grouping an approximate value, which we will call the *empirical mode*, is given by the class mark having the largest frequency.¹ Thus, in Table 17 the empirical mode is 1.745 inches.

12. The Median. Instead of finding the mean, suppose the N variates are arranged in the order of their magnitude. The median is defined as the value which is greater than half the variates and less than the other half. A more precise definition is as follows:

Let x_1, x_2, \dots, x_N be a set of real numbers, which may or may not be all different and suppose they are arranged in order of magnitude so that

$$x_1 \leq x_2 \leq x_3 \leq \dots \leq x_N.$$

Whenever N is odd, $N = 2k - 1$, the median is x_k , the middle one of the x 's. If N is even, $N = 2k$, the median is not uniquely defined unless $x_k = x_{k+1}$, in which case the median is this common value. Otherwise, the definition is satisfied by any value of x belonging to the interval

$$x_k \leq x \leq x_{k+1},$$

and the median is to this extent indeterminate. In this case it is conventional to take

$$\frac{1}{2}(x_k + x_{k+1})$$

as the median.

Example. Find the median of the following set of numbers: 10, 6, 5, 25, 15, 18, 20.

Arranging them in order of magnitude we find the median to be 15 (the mean is 14.14). If we add another value, 37, to make N even, the median is $\frac{1}{2}(15 + 18) = 16.5$ (the mean is 17).

13. Median of a Frequency Distribution. Case I. For a frequency distribution of *continuous* variates, the median is defined as follows:

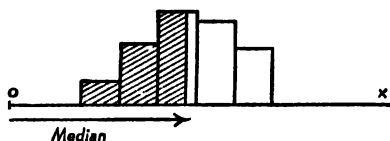
DEFINITION: The median is the value of x for which $\text{cum } f = N/2$.

Given such a frequency distribution we may therefore find its

¹ Another method of computing the mode will be given in a later section.

median by forming a cumulative frequency table and interpolating in the *end-x* column for the value of x corresponding to $N/2$.

The method should be clear from the following illustration.



Find the median for the data of Table 2.

<i>Interval</i>	<i>f</i>	<i>End-x</i>	<i>Cum f</i>
30-39	2	29.5	0
40-49	3	39.5	2
50-59	11	49.5	5
60-69	20	59.5	16
70-79	32	69.5	36
		← Md	← 50
80-89	25	79.5	68
90-99	7	89.5	93
		99.5	100

Here, $N/2 = 50$. This value of *cum f* corresponds to a value of x in the interval 69.5-79.5. Therefore the median is 69.5 plus a fraction of the distance from 69.5 to 79.5. Thus,

$$D_1 \left[\begin{array}{c|c} \text{End-x} & \text{Cum f} \\ \hline d_1 \left[\begin{array}{c} 69.5 \\ \text{Median} \\ 79.5 \end{array} \right] & \left[\begin{array}{c} 36 \\ 50 \\ 68 \end{array} \right] d_2 \end{array} \right] D_2$$

Assuming that the items in any class interval are uniformly distributed over that interval, it follows that the partial differences are proportional to the total differences: $d_1/D_1 = d_2/D_2$. That is,

$$\frac{\text{Median} - 69.5}{79.5 - 69.5} = \frac{50 - 36}{68 - 36}$$

whence,

$$\begin{aligned}\text{Median} &= 69.5 + 10 \left(\frac{14}{32} \right) \\ &= 69.5 + 4.4 = 73.9.\end{aligned}$$

This is called "straight line interpolation" or "interpolation by proportional parts." The reason for these names is made clear in the following diagram.

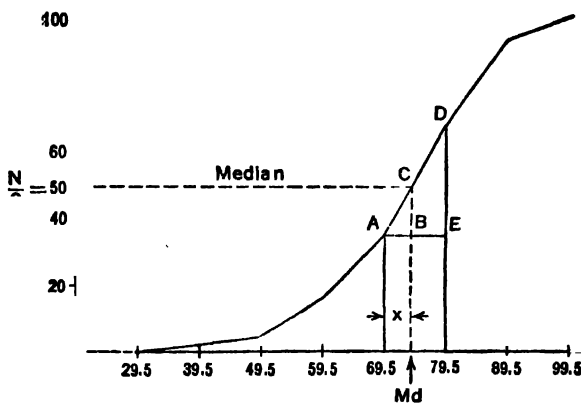


FIG. 8

$\triangle ABC$ is similar to $\triangle AED$

$$\therefore \frac{AB}{AE} = \frac{BC}{ED}$$

$$\begin{aligned}x &= AB = \frac{AE \cdot BC}{ED} \\&= \frac{10(50 - 36)}{68 - 36} \\&= 10 \left(\frac{14}{32} \right) \\&= 4.4\end{aligned}$$

$$\therefore \text{Md} = 69.5 + x = 73.9.$$

The following formula may also be used to compute the median:

$$Md = x_m + \left[\frac{N}{2} - b f_m \right] \frac{c}{f_c}$$

where x_m is the lower end-value of the median class, N is the total frequency, $\sum f_m$ the number of variates below the median class, c the class interval, and f_c the frequency of the median class.

Case II. In the case of a set of discrete variates there may be no value in the set such that the number of variates which are larger than it is equal to the number less than it. Thus in Table 11 the values of x are integers and 35% of the throws yielded 5 or fewer successes and 65% yielded 6 or more successes. Neither $x = 5$ nor $x = 6$, nor any integer, will exactly split the total frequency into two equal parts. Of course a formal application of the definition given in Case I will give a value of x for which $\text{cum } f$ is $N/2$. The difficulty is not so much in the interpretation of the fractional result because the same objection could be cited against the mean. But the real difficulty lies in explaining interpolation in a discontinuous function. We cannot assume that the given frequencies are distributed over the interval from one value of x to the next. Perhaps the best we can do in such cases is to make a statement similar to the one above for Table 11. At least such a statement serves to summarize the situation without artificiality.

14. Graphical Interpretation of Mean, Median, and Mode. The mean corresponds to the abscissa of the point known in mechanics as the centroid of area. If a thin, homogeneous plate of metal cut in

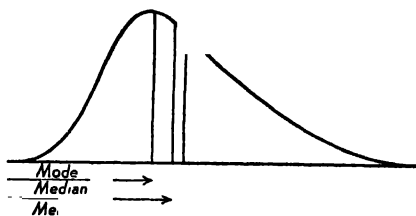


FIG. 9

the shape of a histogram is supported loosely on a horizontal axis through its centroid, the plate will have no tendency to rotate, whatever horizontal direction this axis may assume.

The median of a frequency distribution is the abscissa of a point through which a vertical line will divide the total area of the histogram into two equal parts.

If a distribution could be represented by a smooth curve, then the mode is the abscissa of the highest point on the curve.

Figure 9 shows the position of the three averages in a moderately

skew distribution. If the distribution were perfectly symmetrical then all three of these measures of location would coincide.

There is an interesting empirical relationship between the three quantities which appears to hold for unimodal curves of moderate asymmetry, namely,

$$\text{mean} - \text{mode} = 3(\text{mean} - \text{median}).$$

It is a useful mnemonic to observe that the mean, median, and mode occur in the same order (or reverse order) as in the dictionary; and that the median is nearer to the mean than to the mode, just as the corresponding words are nearer together in the dictionary.¹

15. Discussion. The student primarily interested in the use of these averages in practical statistics might reasonably inquire, "Which of the three averages mentioned should be used in a given problem?" The answer depends upon certain properties peculiar to each average and upon the nature of the data to be averaged.

In most cases the mean is a distinctly superior average. It is rigorously defined, easily computed, and is most tractable in theoretical discussions.

When the median differs considerably from the mean it is likely that the median is the more typical value. The advantage of the median over the mean is recognized in at least three situations:

(a) When occasional and unexpected values occur at the ends of the distribution. In such cases the mean would tend to distort the true representation of the typical value, being unduly influenced by the exceptional values.

(b) When the data are presented in a table left open at one or both ends. For example, suppose the registrar's office of a university reports the following distribution of grades as given in all departments for a semester:

<i>Below 60</i>	60-69	70-79	80-89	90-100
215	1060	2217	1242	506

A *cum f* table may be formed and hence the median can be found without any more information about the values less than 60.

(c) When the observations cannot be measured numerically but can be ordered.

The mode is best adapted to situations where the word "usual" would be appropriate. Unless a large number of items are con-

¹ M. G. Kendall — *The Advanced Theory of Statistics*, vol. I, p. 35. Lippincott.

sidered the mode can have little practical meaning. It is the appropriate average in certain questions of marketing because manufacturers are interested in the type or quality which is usually in demand. Or again, in an investigation concerning wages and cost of living, the mode would reflect the average situation. Also, in a mathematical treatment of frequency curves the concept of the mode is very useful.

Sometimes a distribution has more than one mode, although this is usually due to heterogeneous material. In this course we will be concerned only with unimodal distributions.

The above remarks about the appropriateness of various averages are made from the standpoint of describing and condensing the data *per se*. A few remarks from a different point of view should perhaps be added here. In the theory of sampling, which deals to a large extent with estimating from a sample certain constants in the parent universe, it is shown that the mean has definite advantages. The mean is much more efficient¹ than the median, for example, in estimating the corresponding average in the universe (except in a special case when the universe is an unusual type).

For a more complete treatment of the applicability of these three averages, the student is referred to the following books:

1. *Theory of Statistics* — Yule and Kendall, Ch. VII.
2. *The Mathematics of Statistics* — Burgess, Ch. V.
3. *Mathematical Statistics* — Camp, p. 40.

Exercises

1. State what the empirical mode is in each of Tables 8 to 13.
2. Explain why the median is found from interpolating in the *end-x* column and not the *mid-x* column.
3. Read one or more of the references in §15 and write an essay on the advantages and limitations of the mean, median, and mode.
4. Find the median IQ for the data in Table 7.
5. Find the median for the data in Table 9.

16. Geometric Mean. The geometric mean of a set of N positive values is the N th root of their product. Thus, the geometric mean (G.M.) of two values is the square root of their product, of three values the cube root of their product, and in general for the N values y_1, y_2, \dots, y_N ,

$$(10) \quad \text{G.M.} = [y_1 \cdot y_2 \cdot y_3 \cdots y_N]^{\frac{1}{N}}.$$

¹ See *Economic Control of Manufactured Products* — W. A. Shewhart, p. 280. D. Van Nostrand Co.

Equation (10) lends itself to the use of logarithms and frequently they greatly facilitate the computation of G.M. From (10) we have

$$(11) \quad \log \text{G.M.} = \frac{1}{N} [\log y_1 + \log y_2 + \cdots + \log y_N].$$

Therefore the arithmetic mean of the logarithms of a set of values is the same as the logarithm of the geometric mean of the values themselves.

Examples: Find the geometric mean of

(a) 3, 6, 12, 24, 48.

Solution:

$$\text{G.M.} = [(3^5)(2^{10})]^{1/5} = (3)(2^2) = 12.$$

(b) 7.96, 13.82, 22.95, 35.34.

Solution:

$$\log 7.96 = 0.90091$$

$$\log 13.82 = 1.14051$$

$$\log 22.95 = 1.36078$$

$$\log 35.34 = 1.54827$$

$$\hline 4 \mid 4.95047$$

$$\log \text{G.M.} = 1.23762$$

$$\text{G.M.} = 17.28$$

The geometric mean is the appropriate average when the data are limited at one end of the range and unlimited at the other, and there tends to be a constant rate of change from one y value to the next. This is characteristic of values which tend to form a geometric progression, *i.e.*, which tend to follow the simple exponential law

$$(12) \quad y = ar^x.$$

The student will recall from algebra that a geometric progression can be put in the form

x	0	1	2	\cdots	x
y	a	ar	ar^2	\cdots	ar^x

The value of any term in the y series is a function of the exponent of r since a and r are constants. The functional relationship is therefore represented by (12).

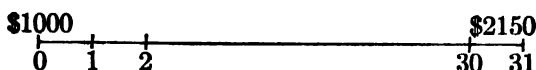
The growth of many quantities in nature follows this law and it is sometimes called the law of natural growth. With x referring to

time, y may represent, for example, the population of a city, the enrollment of a school, the weight of a quantity, or the number of bacteria in a culture. The accumulated amount S , of P dollars invested at i rate of interest, compounded periodically for n periods also takes the form of (12), namely,

$$S = P(1 + i)^n,$$

where r is now $(1 + i)$, a is P , and n and S are the variables corresponding to x and y .

Thus, if \$1000 increased at compound interest to \$2150 in 31 years,



the geometric average rate at which the money increased is found as follows

$$\begin{aligned} r^{31} &= (1 + i)^{31} = \frac{2150}{1000} \\ 1 + i &= (2.15)^{1/31} \\ &= 1.025 \\ i &= 2\frac{1}{2}\%. \end{aligned}$$

Since there was an increase of $\frac{1150}{1000} = 115\%$, the arithmetic average would be $\frac{115}{31} = 3.7\%$ which is also the simple interest rate.

If y in equation (12) represents population, and we are given two values of y corresponding to two dates N years apart, the geometric mean enables us to find a fairer estimate of the value of y at the mid date than would be given by any other average. For example, suppose we are given that the population of a city was 2500 in 1920 and 5000 in 1930. We wish to estimate the population in 1925 and to find the average annual rate of increase. If we are given no other information, our best estimate for 1925 is given by

$$\text{G.M.} = (y_1 \cdot y_2)^{1/2} = (2500 \times 5000)^{1/2} = 3535.$$

The average annual rate of increase is obtained by solving (12) for r as follows:

$$\begin{aligned} 5000 &= 2500r^{10} \\ 2 &= r^{10}. \end{aligned}$$

Hence $r = \sqrt[10]{2} = 1.0718 = 107.18\%$, so that the average annual rate of increase is 7.18%. It is now possible to estimate the population for *any* intermediate year. Thus, for 1928, we have from (12):

$$y = 2500(1.0718)^8 = 4353.$$

The geometric mean is also used in economics in averaging "index numbers" which are essentially the ratios of prices of commodities at one date to their prices at another date. In general it is the appropriate average when emphasis is on the rate or percentage of change rather than the amount.

17. Harmonic Mean. Another average which has long been known and which is required in certain problems is the *harmonic mean* (H.M.). For the N positive values x_1, x_2, \dots, x_N , it is defined as the reciprocal of the arithmetic mean of the reciprocals of the values. In symbols,

$$(13) \quad \text{H.M.} = \frac{1}{\frac{1}{N} \left(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_N} \right)} = \frac{N}{\sum \left(\frac{1}{x_i} \right)}.$$

This measure is used in averaging ratios, such as rates and prices, when certain conditions are agreed upon.

In the case of time rates, we have ratios between two quantities one of which is in units of time, which we will denote by t , and the other is in units of some element like distance or accomplishment or temperature, etc. Denote this second element, different from time, by d . Then we make the following observations:

(a) A rate may be stated either in the form d/t or in the form t/d . Thus, a car which travels at the rate of 30 miles per hour may also be said to travel at the rate of 2 minutes per mile. In this illustration the second form is not the usual way of expressing the rate, but there are cases in which the form t/d is usual. When we say a man takes 10 seconds to run 100 yards we are expressing his rate in time per unit of distance (t/d).

(b) In averaging rates one should first decide whether d or t should properly be the basic or "fixed" element in the discussion. Occasionally there is a difference of opinion about which element should most appropriately be regarded as fixed. For example, suppose a class of students has been given 15 minutes in which to work as many as they can of a given list of problems, and the number of problems

worked correctly by each student recorded. Some educational statisticians would say that *time* should be the fixed element here and that number of problems solved (in a unit of time) should be the variable. Others would say that the number of minutes (t) a student required to work one problem is the proper variable and that a problem (d) should be regarded as the fixed element in the discussion.

In one case the rates are equally weighted in the sense of time and in the other case they are equally weighted in the sense of the element d .

(c) The harmonic mean of the rates expressed in the form d/t gives the same result as the arithmetic mean of the same rates expressed in the form t/d . This is evident from equation (13) if it is written in the form,

$$\frac{1}{\text{H.M.}} = \frac{1}{N} \sum \frac{1}{x_i}$$

and from the fact that rates in one form are merely the reciprocals of the same rates in the other form.

As an illustration, let us consider three cars:

- I { A travels at the rate of 15 miles per hour ($\frac{1}{4}$ mile per minute),
 B travels at the rate of 20 miles per hour ($\frac{1}{3}$ mile per minute),
 C travels at the rate of 30 miles per hour ($\frac{1}{2}$ mile per minute).

But their rates could just as well have been stated as

- II { A travels at the rate of 4 minutes to the mile,
 B travels at the rate of 3 minutes to the mile,
 C travels at the rate of 2 minutes to the mile.

The harmonic mean of the rates as stated in I is 20 miles per hour; i.e., $\frac{2}{3}$ of a mile per minute, and the arithmetic mean of the rates as stated in II is 3 minutes per mile or again, 20 miles per hour. (Verify.)

The third observation, i.e., (c) above, suggests the following discussion. The *arithmetic* mean of the rates in I is $21\frac{2}{3}$ m.p.h. and this is the harmonic mean of the rates as stated in II.

The question arises, which is the correct average, 20 m.p.h. or $21\frac{2}{3}$ m.p.h.? The problem is indeterminate until it is agreed whether time or distance is the fixed element. The correct average will differ

according to the condition agreed upon. This will be made clear in the following analysis.

Case I. Let

$$x_i = \frac{d_i}{t_i}$$

denote the i th rate, $i = 1, 2, \dots, n$. Then the average rate is

$$\frac{D = \text{total distance}}{T = \text{total time}} = \frac{t_1 x_1 + t_2 x_2 + \dots + t_n x_n}{t_1 + t_2 + \dots + t_n}$$

Condition 1. Let distance be the fixed element, i.e., let d be constant. Then $d = t_i x_i$, and $t_i = d/x_i$. Therefore, the expression for average rate becomes

$$\frac{\sum t_i x_i}{\sum \frac{d}{x_i}} = \frac{nd}{d \sum \frac{1}{x_i}} = \frac{1}{\frac{1}{n} \sum \frac{1}{x_i}}$$

which is the harmonic mean.

Condition 2. Suppose t is the fixed element. Then $\sum t_i x_i$ becomes $t \sum x_i$ since t is a constant, and $\sum t$ becomes nt . Hence, we have for the average rate,

$$\frac{D}{T} = \frac{t \sum x_i}{nt} = \frac{1}{n} \sum x_i$$

which is the arithmetic mean.

Case II. Let $x_i = t_i/d_i$ denote the i th rate. Then the average rate is

$$\frac{T = \text{total time}}{D = \text{total distance}} = \frac{\sum t_i}{\sum d_i}$$

Condition 1. Suppose d is the fixed element. Then $t_i = d x_i$ and $d = t_i/x_i$. Hence, we have

$$\frac{T}{D} = \frac{d \sum x_i}{nd} = \frac{\sum x_i}{n}$$

Condition 2. Let t be fixed. Then $d_i = t/x_i$ and the average rate is

$$\frac{T}{D} = \frac{nt}{t \sum \frac{1}{x_i}} = \frac{1}{\frac{1}{n} \sum \frac{1}{x_i}}$$

We therefore state the following rules for averaging rates:

Rule 1. The harmonic mean is used whenever the fixed element is d and the rates are expressed in the form d/t , or when the fixed element is t and the rates are expressed in the form t/d .

Rule 2. The arithmetic mean should be used when the fixed element is t and the rates are expressed in the form d/t , or when the fixed element is d and the rates are expressed in the form t/d .

In the case of prices, which are of course ratios, a similar discussion holds except that now the unit of time is to be replaced by a unit of money. Therefore, prices are ratios between two quantities, one of which is in units of money and the other in units of some commodity or service. They may be stated as so much money per unit of commodity (m/c), or as so many units of commodity per dollar (c/m). Thus, if 100 bushels of wheat are exchanged for 75 dollars of gold, the price of the wheat in terms of gold is $75 \div 100$, or three-fourths of a dollar of gold per bushel of wheat. Contrariwise, the price of gold in terms of wheat is $100 \div 75$, or one and one-third bushels of wheat per dollar of gold. Thus, there are always two prices in any exchange.

The correct average will depend upon how the prices are stated and upon whether a unit of the commodity (or service) or a unit of money is the fixed element.

The following papers in *The Journal of the American Statistical Association* are recommended:

1. "The Nature and Use of the Harmonic Mean" — W. F. Ferger, vol. 26 (1931), pp. 36-40.
2. "Calculating the Geometric Mean from a Large Amount of Data" — Zenon Szatrowski, vol. 41 (1946), pp. 218-220.

Examples

1. In a certain factory a unit of work is completed by A in four minutes, by B in five minutes, by C in six minutes, by D in ten minutes, and by E in twelve minutes. What is their average rate of working? At this rate how many units will they complete in a six-hour day?

Solution. The rates are expressed in the form t/d but it would seem appropriate to regard t as the basic or fixed element since output per unit of time appears to be the important consideration here. So by Rule 1,

$$\text{H.M.} = \frac{5}{\frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{10} + \frac{1}{12}},$$

that is,

$$\text{H.M.} = \frac{300}{48} = 6\frac{1}{4} \text{ minutes per unit.}$$

In 360 minutes they will complete $\frac{4(360)5}{25} = 288$ units.

2. A tourist purchases gasoline at three stations, as follows:

<i>Station</i>	<i>Number of gallons of gasoline for \$1.00</i>
A	5
B	7
C	6

Here the prices are given in the form c/m and it would seem appropriate to regard gallon (c) as the fixed element and prices (m) per gallon as the variable quantities which are to be averaged. Hence, replacing d/t by c/m and "rates" by "prices" in Rule 1, we are led to find the harmonic mean.

$$\begin{aligned} \therefore \text{H.M.} &= \frac{3}{\frac{1}{5} + \frac{1}{6} + \frac{1}{7}} \\ &= \frac{630}{107} \text{ gals. per \$1.00} \\ &= \frac{\$107}{630} \text{ per gal.} \end{aligned}$$

Exercises

- The arithmetic mean of a set of 30 numbers is 82. What is the sum of these numbers?
 - The G.M. of ten numbers is 1.40. What is the product of these ten numbers?
- In chemistry a student was graded 65 in final examination, 85 in recitation and 80 in laboratory. These grades were weighted 1, 2, and 3 respectively. Find the student's average grade.
- At the end of his first semester in college a freshman had credits as follows: 4 hours of mathematics with a grade of 88, 4 hours of English with a grade of 80, 3 hours of history with a grade of 85, and 4 hours of physics with a grade of 78. What was his average grade per hour of credit?
- Find the median of Table 12.
- The population of a city increased in 5 years from 225,000 to 245,000. What was the average increase per year? What was the average annual rate of increase?
- The number of bacteria in a certain culture was found to be 4×10^6 at noon of one day. At noon the next day the number was found to be 9×10^6 . If the number increased at a constant rate per hour, how many bacteria were there at midnight?

7. Find the average (G.M.) rate of interest for five years during which the interest rates were 4.25%, 5.3%, 4.65%, 3.86%, 4.38%.
Hint. $(1+i)^5 = (1.0425)(1.053)(1.0465)(1.0386)(1.0438)$.
8. Find the harmonic mean of the first fifteen positive integers.
9. For two positive numbers, a and b , the geometric mean is $x = \sqrt{ab}$. This is also called the mean proportional between a and b , since $a:x = x:b$. By drawing a semicircle on $a+b$ as diameter, show how the value of x can be constructed geometrically.
10. The following table gives the population of the U. S. at each 10-year census from 1860 to 1920.

<i>Year</i> x	<i>Population</i> (millions)	<i>Ratio of Each Census</i> <i>Figure to Preceding</i>
1860	31.4	
70	38.6	1.23
80	50.2	1.30
90	63.0	1.25
1900	76.0	1.20
10	92.0	1.21
20	105.7	1.15

- What is the average rate of increase per decade? Using this average, estimate the population for 1930 from the 1920 census figure.
11. If a series of positive variates form a geometric progression show that their logarithms form an arithmetic progression.
12. Find the geometric mean of the following:
 (a) 2, 4, 8, 16, 32.
 (b) 47, 92, 123, 218.
13. Given two sets of n positive variates each:

$$x_{11}, x_{12}, x_{13}, \dots, x_{1n}$$

$$x_{21}, x_{22}, x_{23}, \dots, x_{2n}.$$

- Prove that the geometric mean of the ratios of corresponding variates in the two sets is equal to the ratio of their geometric means.
14. (a) For a frequency distribution of positive variates show that (10) becomes

$$\text{G.M.} = \left[x_1^{f_1} \cdot x_2^{f_2} \cdot \dots \cdot x_k^{f_k} \right]^{\frac{1}{N}}$$

where k is the number of different values of x in the set, any exponent f_i is the number of times x_i is repeated, and $N = \sum_1^k f_i$.

- (b) What is the expression for \log G.M. when G.M. is defined as in (a)?
15. A wholesale firm has twelve travelling salesmen who make trips of essentially the same length. Of these, eight make their trip in 20 days and four in 15 days. What is the average time per trip? *Ans.* 18 days.

16. State two rules for averaging prices similar to those given for averaging rates. Give illustrations.
17. Consider any two positive variates x_1 and x_2 . Prove that their geometric mean is equal to the geometric mean between their arithmetic mean and their harmonic mean.
18. (*Burgess*) The following problem arose in a statistical office in Washington during World War I: Suppose 20 boats make 6 trans-atlantic trips each per year, giving as the time for a "turn around" (*i.e.*, time between consecutive departures from the same ports), one-sixth year — approximately 60 days, and that 10 boats make 4 trips per year, giving as their time for a "turn around" one-fourth year, approximately 90 days. (A year of 360 days is used merely for convenience.) What is the average number of days per turn around?

Hint. If we think of the rates expressed as "trips per year" then $x = d/t$. If t is regarded as the fixed element, then by Rule 2 the arithmetic mean is indicated, and $x = 6$ for 20 values of x , and $x = 4$ for 10 values.

If we think of the rates expressed as "days per trip" then $x = t/d$. If t is the fixed element, by Rule 1 the harmonic mean is the correct average, and $x = 60$ for 20 values and $x = 90$ for 10 values. *Ans.* $5\frac{1}{2}$ trips per year or 67.5 days per trip.

19. Show that if $2a$ is the harmonic mean of the two rational numbers b and c , then the sum of the squares of the three numbers a , b , and c is the square of a rational number.
- (Reference: *American Mathematical Monthly*, June 1935, p. 394.)
20. (a) If A , G , and H represent, respectively, the arithmetic, geometric, and harmonic means of N unequal positive variates, prove that

$$H < G < A$$

(Reference: *Burgess' text*, p. 101.)

- (b) What can you say if the N positive variates are equal?
21. A plane travels one half of a given distance D in miles at a speed of x_1 miles per hour, and the remaining half distance at a speed of x_2 miles per hour. Show that the average speed for the entire distance is the harmonic mean of x_1 and x_2 . Half of this average speed is called the "radius of action per hour"; *i.e.*, it is the outbound distance that a plane can travel and return in one hour. The "radius of action" of a plane would be the "radius of action per hour" multiplied by the number of hours in flight.

CHAPTER IV

MOMENTS

1. Moments about an Arbitrary Origin. One of the general problems of statistics is to summarize and characterize data. In the words of R. A. Fisher,

A quantity of data which by its mere bulk may be incapable of entering the mind is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.¹

These "relatively few quantities" are usually expressed in terms of moments. Moments are of different orders and the student is already familiar with what is now to be known as the first moment, namely, the arithmetic mean of the first powers of the variates. We will also need in our work the arithmetic means, respectively, of the second, third, and fourth powers of the variates. With reference to an arbitrary origin, moments are denoted by ν (the Greek letter nu) with a subscript specifying the order.

The first four moments, relative to the x -origin and in the x unit, are defined as follows:

$$(1) \quad \left\{ \begin{array}{l} \nu_1 = \frac{1}{N} \sum f_i x_i = \bar{x} \\ \nu_2 = \frac{1}{N} \sum f_i x_i^2 \\ \nu_3 = \frac{1}{N} \sum f_i x_i^3 \\ \nu_4 = \frac{1}{N} \sum f_i x_i^4, \end{array} \right.$$

i varying from 1 to k .

A more general definition of the ν 's is

$$(1a) \quad \nu_r = \frac{1}{N} \sum_1^k f_i (x_i - x_0)^r$$

¹ *Foundations of Theoretical Statistics*, Philosophical Transactions of the Royal Society, vol. 222A (1922), p. 309.

for the r th moment about an arbitrary point x_0 . When $x_0 = 0$ and $r = 1, 2, 3$, and 4 , we have the definitions stated in (1). If $r = 0$ we have the zeroth moment and $\nu_0 = 1$.

In statistics we work with moments per unit frequency. The term "moment" has its origin in mechanics where we speak of the "moment of a force." Suppose we have a rigid bar, called a lever, with one point of support known as a fulcrum (Figure 10). If a force f_1 is applied to the lever at a distance x_1 from the fulcrum O , the product $x_1 f_1$ is called the moment of the force. If there are two or more such forces f_1, f_2, \dots, f_k , acting in the same direction, and at the distances x_1, x_2, \dots, x_k , respectively from O , the total moment of all these forces is

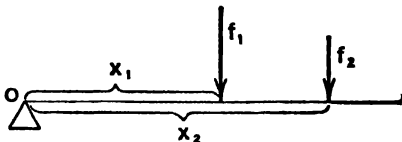


FIG. 10

$$f_1 x_1 + f_2 x_2 + \dots + f_k x_k = \sum f_i x_i.$$

If the distances x are squared, we have $\sum f_i x_i^2$ as the total second moment, and $\sum f_i x_i^r$ represents the r th moment.

It is by analogy with this mechanical concept that the expressions in (1) are called statistical moments (per unit frequency) about zero as origin.

Exercises

- Write out the expanded form of the ν 's defined in (1).
- Calculate the values of ν_1, ν_2 , and ν_3 for the following distributions:

(a)

(b)

x	f	x	f
0	1	-3	1
1	3	-2	3
2	5	-1	5
3	10	0	5
4	5	1	3
5	2	2	1

- Prove that ν_0 is always equal to unity.
 - Prove that moments of even order are always positive or zero, but that moments of odd order may be positive, negative, or zero.

(c) Show that the odd moments are all zero if both the x 's and f 's are symmetrical with respect to the origin of x , as, for example,

x	-1.5	-1.0	-0.5	0.5	1.0	1.5
f	1	2	3	3	2	1

2. Moments in Units of the Class Interval. In Chapter III, §8, the mean in the x unit was obtained by first finding the mean in the u unit, viz., $\frac{1}{N} \sum f_i u_i$, and then changing over into the x unit by multiplying by the interval c . In our subsequent work, which requires the higher moments, we shall find it convenient to use a similar procedure, and find those moments in the u unit, where $u = (x - x_0)/c$. It is desirable, therefore, in labeling the moments for any distribution, to specify whether they are in the unit of x or u . This is commonly done by the use of a second subscript on ν . Thus $\nu_{r,u}$ denotes the r th moment in the u unit and relative to the u -origin. Therefore,

$$(2) \quad \left\{ \begin{array}{l} \nu_{1;u} = \frac{1}{N} \sum f_i u_i = \bar{u} \\ \nu_{2;u} = \frac{1}{N} \sum f_i u_i^2 \\ \nu_{3;u} = \frac{1}{N} \sum f_i u_i^3 \\ \nu_{4;u} = \frac{1}{N} \sum f_i u_i^4 \end{array} \right.$$

Similarly, $\nu_{r;x}$ will mean $\frac{1}{N} \sum f_i x_i^r$. When there is no ambiguity, the second subscript on ν may be omitted.

3. Moments about the Mean. Formulas (1) and (2) define the moments taken about zero as origin although in different units. When the mean is chosen as origin we have the most important set of moments in the theory of statistics. In this case the Greek letter μ (mu) is used to denote the moments, and it is always understood that the use of μ specifies the mean as origin. It does not, however, designate the unit, so the second subscript may still be necessary. Therefore, the r th moment about the mean is defined by either of the

following expressions:

$$(3) \quad \left\{ \begin{array}{l} \mu_{r,x} = \frac{1}{N} \sum f_i (x_i - \bar{x})^r \\ \mu_{r,u} = \frac{1}{N} \sum f_i (u_i - \bar{u})^r. \end{array} \right.$$

The mean is a sort of balance point. If weights proportional to the frequencies are suspended along a horizontal bar at distances from one end proportional to the numbers representing the class marks, then the bar will balance at the weighted mean of the distances. In mechanics this point is known as the abscissa of the *center of gravity* or *centroid*. Theorem VI of Chapter III, §7, is another way of saying that the given distribution is in equilibrium about this point.

4. Relations between the μ 's and ν 's. We shall see that the descriptive constants mentioned at the beginning of the chapter are defined in terms of the moments about the mean, but the moments about an arbitrary point are easier to calculate. In other words, what we desire are the values of μ_r , but their computation directly from the definitions (3) may be very laborious even in the u unit due to the fact that $(u - \bar{u})$ usually involves decimals. Raising these decimals to the second, third, and fourth powers becomes tedious even with the aid of a computing machine. On the other hand, the ν 's defined in (2) are readily computed. Therefore, instead of computing the μ 's directly we obtain them indirectly from the ν 's. The relations between the μ 's and ν 's can be found by expanding, by the Binomial Theorem, either of the expressions following the \sum 's in (3) for $r = 2, 3, 4$. This is done in the u unit as follows:

$$\begin{aligned} \mu_2 &= \frac{1}{N} \sum f_i (u_i - \bar{u})^2 \\ &= \frac{1}{N} \sum f_i u_i^2 - 2\bar{u} \cdot \frac{1}{N} \sum f_i u_i + \bar{u}^2 \\ &= \nu_2 - 2\bar{u}\nu_1 + \bar{u}^2 \\ (4) \quad &= \nu_2 - (\nu_1)^2, \text{ since } \bar{u} = \nu_1 \\ \mu_3 &= \frac{1}{N} \sum f_i (u_i - \bar{u})^3 \\ (5) \quad &= \nu_3 - 3\nu_2 \cdot \nu_1 + 2(\nu_1)^3 \\ (6) \quad \mu_4 &= \nu_4 - 4\nu_3 \cdot \nu_1 + 6\nu_2(\nu_1)^2 - 3(\nu_1)^4. \end{aligned}$$

These formulas are important and the student should be able to derive them. It should be apparent that these moment relations hold also in the x unit. However, if we have the μ 's in the u unit and we desire them in the x unit they may be found as follows:

$$(7) \quad \begin{cases} \mu_{2;x} = c^2 \mu_{2;u} \\ \mu_{3;x} = c^3 \mu_{3;u} \\ \mu_{4;x} = c^4 \mu_{4;u} \end{cases}$$

The first of the relations given in (7) is proved below. The others may be proved in a similar manner.

$$\begin{aligned} \mu_{2;x} &= \frac{1}{N} \sum f_i (x_i - \bar{x})^2 \text{ by definition,} \\ &= \frac{1}{N} \sum f_i (x_0 + cu_i - x_0 - c\bar{u})^2 \text{ by (4a) and (5), Chapter III,} \\ &= \frac{c^2}{N} \sum f_i (u_i - \bar{u})^2 = c^2 \mu_{2;u}. \end{aligned}$$

We see that the indirect method of computing the μ 's (in the u unit) involves two steps. First the ν 's are computed according to the definitions in (2). This step is illustrated in Table 18. Then we calculate the μ 's by substituting the computed ν 's in relations (4), (5), and (6). The μ 's in the x unit could then be obtained, if desired, by means of (7).

Before proceeding with the second step it is desirable to check the ν 's or, at least, the totals of the columns from which they are obtained. This can be done if we have another column headed $f(u+1)^4$, and observe that

$$\sum f(u+1)^4 = \sum fu^4 + 4\sum fu^3 + 6\sum fu^2 + 4\sum fu + \sum f.$$

This is known as *Charlier's check*. An alternative one is to check the entries in the column fu^4 against the proper entries in Pearson's *Tables for Statisticians and Biometricians*, Table L.

Charlier's check is a necessary but not a sufficient check. That is to say, compensating errors may occur which this check would not detect. However, the occurrence of such errors is very unlikely.

Applying Charlier's check to Table 18 we have

$$1220 = 1088 + 4(-236) + 6(176) + 4(-20) + 100 = 1220.$$

TABLE 18 — MOMENTS FOR DISTRIBUTION OF GRADES

Data		Computations					
x	f	u	fu	fu^2	fu^3	fu^4	$f(u + 1)^4$
34.5	2	-4	-8	32	-128	512	162
44.5	3	-3	-9	27	-81	243	48
54.5	11	-2	-22	44	-88	176	11
64.5	20	-1	-20	20	-20	20	0
74.5	32	0	0	0	0	0	32
84.5	25	1	25	25	25	25	400
94.5	7	2	14	28	56	112	567
Sums	100		-20	176	-236	1088	1220
$\frac{1}{N}$ Sums	1		- .20 $\nu_{1;u}$	1.76 $\nu_{2;u}$	-2.36 $\nu_{3;u}$	10.88 $\nu_{4;u}$	For Charlier's check

Hence we may proceed with confidence to compute the μ 's. Using relations (4), (5), and (6):

$$\mu_{2;u} = 1.76 - (-.20)^2 = 1.72$$

$$\mu_{3;u} = -2.36 - 3(1.76)(-.20) + 2(-.20)^3 = -1.320$$

$$\begin{aligned}\mu_{4;u} &= 10.88 - 4(-2.36)(-.20) + 6(1.76)(-.20)^2 - 3(-.20)^4 \\ &= 9.4096.\end{aligned}$$

The following check, which can be handled readily on a machine, may be used to check the μ 's:

$$\begin{aligned}\nu_4 &= \frac{1}{N} \sum f_i x_i^4 = \frac{1}{N} \sum f_i [(x_i - \nu_1) + \nu_1]^4 \\ &= \mu_4 + 4\mu_3\nu_1 + 6\mu_2\nu_1^2 + \nu_1^4.\end{aligned}$$

Before explaining the applications of μ_2 , μ_3 , and μ_4 we present some exercises which will aid the student in mastering the procedure thus far developed.

Exercises

- (a) Verify relations (4), (5), and (6).
 (b) Show that these relations hold also in the x unit.
 (c) Prove that $\mu_1 = 0$ in any unit.
 (d) When $f = 1$, show that

$$\mu_{2;x} = \frac{1}{N} \sum_{i=1}^N x_i^2 - \bar{x}^2$$

2. Verify the relations given in (7).
3. Using Table 18 as a model find the ν 's for Iowa City rainfall by extending Table 17.
4. Find the μ 's from your results in Exercise 3 above.

5. Standard Deviation. Formula (4), $\mu_2 = \nu_2 - \nu_1^2$, is perhaps the most important of the moment relations for elementary statistics. It states that the second moment about the mean is equal to the second moment about zero diminished by the square of the mean measured from zero.

Many of the definitions in statistics are essentially those of physics and mechanics. The analogy between the mean and centroid has been mentioned. The above statement about formula (4) is a well-known proposition in mechanics when the word centroid is substituted for mean.

In mechanics the equivalent of $N\mu_2$ is called the *moment of inertia* (about the axis through the centroid) and $(\mu_2)^{1/2}$ is the *radius of gyration*. These notions are carried over in statistics. Suppose a thin metal plate in the shape of a histogram is rotating about a vertical axis through its centroid. There is a distance from the centroid at which the entire mass of the histogram could be concentrated without changing its moment of inertia. This distance is the square root of μ_2 . It is an *average rotational radius* for all particles of the rotating mass. In statistics, $(\mu_2)^{1/2}$ is called the *standard deviation* and is denoted by the small Greek letter σ . Therefore we have

$$(8) \quad \begin{cases} \sigma_x = \sqrt{\mu_{2:x}} \\ \sigma_u = \sqrt{\mu_{2:u}} \\ \sigma_z = c\sigma_u. \end{cases}$$

We shall see later that σ is a measure of what is called *dispersion*. More precisely, it measures the extent to which the data are spread out "on the average" on either side of the mean. (See Figure 11.) The student will obtain a more complete understanding of σ as the course develops.

The mean and standard deviation are always expressed finally in the same units as the variates. If x represents inches, we desire the mean and standard deviation in inches. When obtained they should be labelled appropriately.

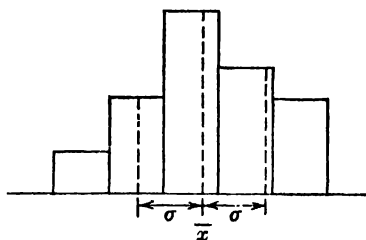


FIG. 11

Example. For Table 18, we have

$$\bar{x} = c\bar{u} + x_0 = 10(-.20) + 74.5 = 72.5\%$$

$$\sigma_u = \sqrt{\mu_{2:u}} = (1.72)^{1/2} = 1.31$$

$$\sigma_x = c\sigma_u = 10(1.31) = 13.1\%$$

Thus, we have explained the use of the first and second moments.

The student will observe that the change from σ_u to σ_x does not involve x_0 . The standard deviation is affected by the change in units but is independent of the origin of reference. To prove this let $x' = x - x_0$, whence $\bar{x}' = \bar{x} - x_0$ (why?). Then

$$\begin{aligned}\sigma_{x'}^2 &= \mu_{2:x'} = \frac{1}{N} \sum f_i (x_i' - \bar{x}')^2 \\ &= \frac{1}{N} \sum f_i [x_i - x_0 - \bar{x} + x_0]^2 \\ &= \frac{1}{N} \sum f_i (x_i - \bar{x})^2 \\ &= \mu_{2:x} = \sigma_x^2.\end{aligned}$$

This suggests the more general

Theorem. *The value of μ_r remains invariant under a transformation which changes only the origin of reference of the variates.*

The student is asked to prove the equivalent of this theorem in Exercise 3 after §9.

6. Standard Units. The above section explains μ_2 . There remains the explanation of μ_3 and μ_4 . We will lead up to this by defining standard units. We have mentioned the transformation $x' = x - \bar{x}$. Another very useful transformation consists in measuring such deviations from the mean in units of the standard deviation,

σ_x , of the entire distribution. They are then known as standard units and will be designated by t . Thus,

$$(9) \quad t = \frac{x - \bar{x}}{\sigma_x} = \frac{x'}{\sigma_x}.$$

Graphically, this translates the origin to the mean and measures distances along the horizontal axis in terms of σ_x . It is a special case of the more general transformation

$$u = \frac{x - x_0}{c}.$$

The significant characteristic of the t variate is its independence of the unit in which the original measurements were taken. For example, suppose we were concerned with obtaining the linear measurements of a set of individuals. One distribution of variates would result if the measurements were made in feet. In this case x' , \bar{x} , and σ_x would also be in feet. If the measurements were taken in inches, then x' , \bar{x} , and σ_x would be in inches, and each of these values would be, numerically, twelve times as large as the corresponding numbers in the first distribution. However, the variates expressed in standard units would be the same for the two distributions. Thus if

$$\bar{x} = 50 \text{ ft.} = 50(12) \text{ in.},$$

and

$$\sigma_x = 5 \text{ ft.} = 5(12) \text{ in.},$$

then for an individual measurement of $x = 60 \text{ ft.} = 60(12) \text{ in.}$, we have

$$t = \frac{10 \text{ ft.}}{5 \text{ ft.}} = \frac{10(12) \text{ in.}}{5(12) \text{ in.}}$$

$$t = 2 = 2.$$

It is obvious, therefore, that standard units provide a basis for comparing distributions. Moreover, they make possible important simplifications in certain mathematical operations.

With the aid of a computing machine, a distribution may be easily transformed into standard units by means of the so-called continuous process. To illustrate, suppose for the distribution of Table 9 (§11, Chapter I), it has been found that

$$\bar{x} = 47.712 \text{ lbs.}$$

$$\sigma_x = 5.772 \text{ lbs.}$$

By relation (9), then,

$$t = \frac{x - 47.712}{5.772} = .17325x - 8.2661.$$

Referring to the discussion of the continuous method given in the Introduction, we observe that here $k = -8.2661$, $n = .17325$, and we desire the values of t corresponding to the values of x given in Table 9. For the values of x such that $nx < k$, we write the above relation in the form

$$-t = 8.2661 - .17325x.$$

The procedure¹ now is to register 8.266100 on the product register, punch the constant factor .17325 on the keyboard, and then by turning the crank backward so that the successive values of x appear on the revolution register, we subtract from k the products of this multiplier and the values of x . The various values of x are built over from one to another without clearing the dial. The resulting values of $-t$ are read at each stage from the product register until we get $-t = 0.383$. From here, $nx > k$, so we clear the dials and start over using the original form of the relation between x and t . We now register -8.266100 on the product register by turning the crank backward, punch .17325 on the keyboard, and turn the crank forward to form the values of x on the revolution register. The values of t are read as before from the product register at each stage of the build-over process. In this way the following set of standard variates is obtained:

TABLE 19

x	f	t
29.5	1	-3.155
33.5	14	-2.462
37.5	56	-1.770
41.5	172	-1.076
45.5	245	-0.383
49.5	263	0.310
53.5	156	1.003
57.5	67	1.696
61.5	23	2.389
65.5	3	3.082

¹ If automatic machines are available the instructor will explain the procedure.

We see from Table 19 that a range of $t = \pm 3$ takes in practically all the variates. This is typical of the more common distributions.

If $\bar{x} = 0$, then $t = x/\sigma$ and the origin of t is the same as the origin of x . Some writers use X to denote the variates (*i.e.*, pounds, dollars, temperatures, etc.), and use x to denote deviations from the mean. In that notation, $t = x/\sigma$ would have the same meaning as our equation (9). Occasionally in later chapters we shall find it convenient to designate deviations from the mean by x (instead of x'). If so, it will be stated that the origin of x is at the mean or centroid.

7. Moments in Standard Units. The moments in standard units are denoted by the Greek letter alpha, α . Thus for the r th moment in standard units, we have $\alpha_r = \frac{1}{N} \sum f_i t_i^r$. However, it is not necessary to transform the variates into t units in order to compute the α 's. We shall show that they are functions of the μ 's. Thus

$$\begin{aligned}\alpha_r &= \frac{1}{N} \sum f_i t_i^r && \text{by definition} \\ &= \frac{1}{N} \sum f_i \left(\frac{x_i - \bar{x}}{\sigma_x} \right)^r && \text{from (9)} \\ &= \frac{1}{\sigma_x^r} \frac{1}{N} \sum f_i (x_i - \bar{x})^r. && \text{Why?}\end{aligned}$$

Hence

$$\begin{aligned}(10) \quad \alpha_r &= \frac{\mu_{r:x}}{(\sigma_x)^r} && \text{Why?} \\ &= \frac{\mu_{r:x}}{(\mu_{2:x})^{r/2}} && \text{from (8).}\end{aligned}$$

Letting $r = 1, 2, 3, 4$ in (10) we have

$$(10a) \quad \left\{ \begin{aligned} \alpha_1 &= \frac{\mu_{1:x}}{\sigma_x} = 0 \\ \alpha_2 &= \frac{\mu_{2:x}}{\sigma_x^2} = 1 \\ \alpha_3 &= \frac{\mu_{3:x}}{(\sigma_x)^3} \\ \alpha_4 &= \frac{\mu_{4:x}}{(\sigma_x)^4} \end{aligned} \right.$$

It is obvious that α_1 and α_2 are abstract numbers. This is also the case for the other α 's. In the expressions for α_3 and α_4 both numera-

tor and denominator are of the same dimension. That is to say, in $\alpha_3 = \mu_3/\sigma^3$ both numerator and denominator are the cubes of whatever unit is used in the original measurements, and therefore their ratio is of zero dimension, a pure number. Similarly, in $\alpha_4 = \mu_4/\sigma^4$ both numerator and denominator are the four powers of the same unit, and therefore α_4 is an abstract number.

Some writers use g_1 instead of α_3 and g_2 for $\alpha_4 - 3$.

8. Use of α_3 and α_4 . Since α_1 and α_2 have the same values for all frequency distributions, their computation contributes nothing to the description or characterization of a distribution. But the values of α_3 and α_4 depend upon the shape of the histogram representing a distribution, and are therefore useful in distinguishing between types of distributions. Thus, we observe that

$$\mu_3 = \frac{1}{N} \sum f(x - \bar{x})^3$$

is a measure of asymmetry about the mean. If the variates are distributed symmetrically about \bar{x} then $\mu_3 = 0$. But if the positive deviations from the mean outweigh the negative deviations then $\mu_3 > 0$, whereas if the negative deviations predominate, then $\mu_3 < 0$. Cubing the deviations gives a measure which is sensitive both to their size and sign but the result is in cubic units. Now symmetry, or lack of it, is not a function of the original units of measurement, so if we divide μ_3 by σ^3 we get a pure number. Thus α_3 is a satisfactory measure for comparing symmetry in distributions of different units of measurement.

The quantity α_4 measures a characteristic called "kurtosis." It refers to the relative number of variates in the vicinity of the mean. More will be said about α_3 and α_4 later on. At this time, emphasis should be placed upon their calculation rather than upon the information which they yield.

Inasmuch as the α 's are independent of the unit of measurement, they may be computed from the moments in the u unit. Changing these moments into the x unit would only introduce the same factor into the numerator and denominator, which would of course divide out. Thus:

$$\alpha_3 = \frac{\mu_{3:x}}{\sigma_x^3} = \frac{C^3 \mu_{3:u}}{C^3 \sigma_u^3} = \frac{\mu_{3:u}}{\sigma_u^3},$$

$$\alpha_4 = \frac{\mu_{4:x}}{\sigma_x^4} = \frac{C^4 \mu_{4:u}}{C^4 \sigma_u^4} = \frac{\mu_{4:u}}{\sigma_u^4}.$$

For Table 18 we have

$$\alpha_3 = \frac{-1.320}{(1.72)(1.31)} = -0.586,$$

$$\alpha_4 = \frac{9.4096}{(1.72)^2} = 3.18.$$

Although no limits can be placed on the possible values which α_3 and α_4 may take, it may be said that for the more common distributions α_4 fluctuates around 3 and α_3 is usually not more than 2 nor less than -2. We cannot go into the theoretical reasons for these values and we mention them here merely to guide the student as to what is a reasonable result to expect in the exercises in this book. In this connection, the inequality¹

$$\alpha_4 \geq \alpha_3^2 + 1$$

may also prove useful. When the numerical value of α_3 is large, the distribution may be of the J-shaped type which is an extreme form of the asymmetrical type. However, these types cannot always be distinguished by elementary methods if the original data are not available.

9. Summary. The quantities \bar{x} , σ_x , α_3 , and α_4 are called the descriptive constants of the distribution. They (together with N) are the "relatively few quantities" (§1) which, in certain cases, contain all the relevant information in the distribution. Table 20 will serve as a model for the procedure which the student should follow in computing these quantities. Of course, if the work is done on a computing machine, only the totals of the power sums need be recorded. The detail of the columns may be omitted. In Table 20, $c = 1$, so $\sigma_x = \sigma_u$. Obviously, this would not be true in general.

The calculation of the ν 's proceeds naturally as an extension of the work required to compute \bar{x} for a frequency distribution. Thus to obtain \bar{x} we first compute $\nu_{1:u}$ and then obtain \bar{x} from the relation

$$\bar{x} = c\bar{u} + x_0.$$

To obtain the standard deviation we need the value of ν_2 because σ_x is found from the relations

$$\mu_2 = \nu_2 - \bar{u}^2$$

$$\sigma_u = \sqrt{\mu_{2:u}}$$

$$\sigma_x = c\sigma_u.$$

¹ "A Note on Skewness and Kurtosis" — J. Ernest Wilkins, Jr. *Annals of Mathematical Statistics*, vol. 15 (1944), pp. 333-335.

The next chapter is devoted to a discussion of dispersion of which σ_x is a measure. To be sure, the standard deviation is only one of several measures of dispersion, just as the mean is only one of several averages. But both the mean and the standard deviation play important roles in the theory and practice of statistics. It is important to master the pattern by which they are computed in a frequency distribution.

In order to compute α_3 and α_4 we first require ν_3 and ν_4 (in addition to ν_1 and ν_2). Then μ_3 and μ_4 are obtained from (5) and (6). Finally,

$$\mu_r = \frac{\mu_{r,u}}{(\sigma_u)^r}$$

is computed for $r = 3$ and $r = 4$. The characteristics of a distribution which α_3 and α_4 describe will be discussed in Chapter VI and again in Part II. In elementary work they are less important than \bar{x} and σ_x .

With regard to the number of decimal places to be retained in computations, the author agrees with Dr. Shewhart who says: "It does not appear feasible . . . to lay down simple, practical, and infallible rules." Reasons in support of this opinion are stated in his book,¹ pp. 79-80. For other remarks in this connection, the reader is referred to the books by Walker and Scarborough which are cited in our Introduction.

Exercises

- (a) What is the numerical value of the mean of any distribution of variates expressed in t units?
(b) What is the standard deviation of such a distribution? *Hint:* $\sigma_t = \sqrt{\alpha_2}$.
- (a) Show that $(x - \bar{x}) = c(u - \bar{u})$ and hence that $t = (u - \bar{u})/\sigma_u$.
(b) Show that we obtain the same results for the α 's if we take

$$t = \frac{u - \bar{u}}{\sigma_u}.$$

- Prove: If any constant is added algebraically to each variate of a series the values of μ_r for the new series will be identical with the corresponding values of μ_r of the original series.
- Suppose each variate is multiplied by a constant. What effect would this have on \bar{x} , σ_x , α_3 , and α_4 ?

¹ See footnote, p. 52.

5. Show that the standard deviation of x may be written

$$\begin{aligned}\sigma_x &= \left[\frac{1}{N} \sum f_i (x_i - \bar{x})^2 \right]^{1/2} \\ &= \left[\frac{1}{N} \sum f_i x_i^2 - \bar{x}^2 \right]^{1/2}.\end{aligned}$$

6. Prove the general relation

$$\mu_{r;x} = c^r \mu_{r;u}$$

of which the relations given in (7) are special cases when $r = 2, 3, 4$.

Hint: $(x - \bar{x}) = c(u - \bar{u})$.

7. (a) Show that $\alpha_0 = 1$.
 (b) Show that $\sigma^r = (\mu_2)^{r/2}$ in both the x and u units.
8. Prove from (4) that μ_2 is less than or at most equal to ν_2 , the same unit being used in each case.
9. Find \bar{x} , σ_x , α_3 , and α_4 for Iowa City rainfall using your results from Problem 4 of the preceding set of Exercises.

Ans.

$$\bar{x} = 2.80 \text{ in.}$$

$$\alpha_3 = 1.29,$$

$$\sigma_x = 2.01 \text{ in.}$$

$$\alpha_4 = 4.58.$$

10. Using Table 20 as a model find \bar{x} , σ_x , α_3 , and α_4 for the distributions in §11, Chapter I, according to the direction of the instructor.

TABLE 20 — SPECIMEN WORKSHEETS FOR COMPUTING THE CHARACTERIZING CONSTANTS OF A DISTRIBUTION

Subject: Span among Adult Males (Table 13)

x	f	u	uf	u^2f	u^3f	u^4f	$(u + 1)^4f$
58.5	1	-11	- 11	121	-1,331	14,641	10,000
59.5	2	-10	- 20	200	-2,000	20,000	13,122
60.5	1	- 9	- 9	81	- 729	6,561	4,096
61.5	6	- 8	- 48	384	-3,072	24,576	14,406
62.5	7	- 7	- 49	343	-2,401	16,807	9,072
63.5	22	- 6	-132	792	-4,752	28,512	13,750
64.5	55	- 5	-275	1,375	-6,875	34,375	14,080
65.5	111	- 4	-444	1,776	-7,104	28,416	8,991
66.5	146	- 3	-438	1,341	-3,942	11,826	2,336
67.5	182	- 2	-364	728	-1,456	2,912	182
68.5	229	- 1	-229	229	- 229	229	0
69.5	265	0	0	0	0	0	265
70.5	263	1	263	263	263	263	4,208
71.5	217	2	434	868	1,736	3,472	17,577
72.5	176	3	528	1,584	4,752	14,256	45,056
73.5	132	4	528	2,112	8,448	33,792	82,500
74.5	82	5	410	2,050	10,250	51,250	106,272
75.5	48	6	288	1,728	10,368	62,208	115,248
76.5	20	7	140	980	6,860	48,020	81,920
77.5	16	8	128	1,024	8,192	65,536	104,976
78.5	12	9	108	972	8,748	78,732	120,000
79.5	3	10	30	300	3,000	30,000	43,923
80.5	1	11	11	121	1,331	14,641	20,736
81.5	2	12	24	288	3,456	41,472	57,122
82.5	1	13	13	169	2,197	28,561	38,416
Sums	2,000		886	19,802	35,710	661,058	928,254
(Sums)/ N			.443	9.901	17.855	330.529	
			\bar{u}	ν_2	ν_3	ν_4	

Charlier's check:

$$\sum (u + 1)^4 f = \sum u^4 f + 4 \sum u^3 f + 6 \sum u^2 f + 4 \sum u f + \sum f$$

$$928,254 = 661,058 + 4(35,710) + 6(19,802) + 4(886) + 2,000 = 928,254$$

Computations:

$$\bar{x} = c\bar{u} + x_0 = (1)(.443) + 69.5 = 69.943 \text{ in.}$$

$$\bar{u}^2 = .196249$$

$$\bar{u}^3 = .086938, \bar{u}^4 = .038514$$

$$\mu_2 = \nu_2 - \bar{u}^2$$

$$= 9.901 - .196249 = 9.704751$$

$$\sigma_u = \sqrt{9.704751} = 3.115$$

$$\sigma_x = c\sigma_u = (1)(3.115) = 3.115 \text{ in.}$$

$$\mu_3 = \nu_3 - 3\nu_2\bar{u} + 2\bar{u}^3$$

$$= 17.855 - 3(9.901)(.443) + 2(.086938)$$

$$= 17.855 - 13.158429 + .173876$$

$$= 4.870447$$

$$\mu_4 = \nu_4 - 4\nu_2\bar{u}^2 + 6\nu_3\bar{u} - 3\bar{u}^4$$

$$= (330.529) - 4(17.855)(.443) + 6(9.901)(.196249) - 3(.028514)$$

$$= 330.529 - 31.639060 + 11.658368 - .115542$$

$$= 310.432769$$

$$\sigma_u^3 = (3.115)(9.704751) = 30.230299$$

$$\sigma_u^4 = (9.704751)^2 = 94.182192$$

$$\alpha_3 = \frac{4.870447}{30.230299} = .161$$

$$\alpha_4 = \frac{310.432766}{94.182192} = 3.296$$

Summary:

$$\bar{x} = 69.943 \text{ in.};$$

$$\alpha_3 = 0.161;$$

$$\sigma_x = 3.115 \text{ in.};$$

$$\alpha_4 = 3.296.$$

10. Sheppard's Corrections. The moments of a frequency distribution are computed on the assumption that each variate value in a class interval has the value of the class mark for that interval. This has the effect of replacing the actual data by somewhat fictitious data assigned arbitrarily at the central values of the intervals. Evidently a very coarse grouping might be misleading and it can be shown mathematically that the above assumption introduces a systematic error, called a grouping error, in the results obtained for the second and fourth moments about the mean but does not affect μ_1 and μ_3 . To eliminate this systematic tendency certain corrections are applied to μ_2 and μ_4 .

The derivation of these corrections is beyond the scope of an elementary course, but it may be worth while to see why it is that corrections are necessary for some moments and not for others. The following argument is intended only as a pedagogical device to give a plausible explanation. Suppose a smooth curve represents the

true frequency distribution while the histogram represents the distribution with class marks as the variates. Since the moments are computed from the distribution represented by the histogram, we scarcely expect our results to be exactly the values of the moments of the true distribution, which are, of course, what we seek. In using the distribution represented by the histogram, we are neglecting, for each rectangle, the little area under the curve shaded *A* and substituting for it the little area shaded *B*. Suppose that, in general, *B* is a little larger than *A*, as shown in Figure 12. The excess of *B*

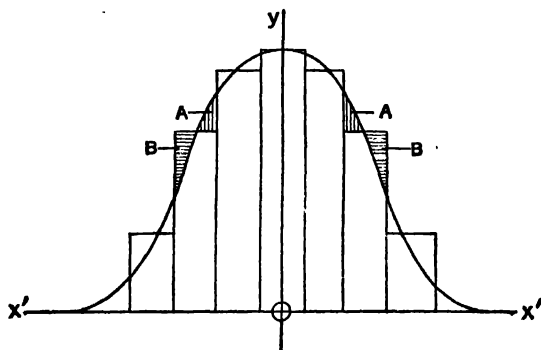


FIG. 12

over *A* for those rectangles to the left of \bar{x} will be negative; the corresponding excess for those rectangles to the right of \bar{x} will be positive. This may be readily understood by considering these little areas as approximate triangles whose bases are negative or positive according as they are to the left or right of \bar{x} . These excesses for all the rectangles, both positive and negative, are involved in taking the summation $\sum f_i(x_i - \bar{x})^r$ for the moments. When *r* is an odd number, as 1 or 3, the excesses show up with their algebraic signs and therefore, over the range of the distribution, the positive excesses just about offset the negatives ones. But in the case of the even moments, all the excesses now become positive so that the errors accumulate and the final results for these moments are too large.

To reduce these errors due to grouping, W. F. Sheppard has demonstrated¹ that the following corrections should be applied. It should

¹ Students familiar with more advanced mathematics will find an interesting discussion of systematic errors and references to papers dealing with Sheppard's corrections in an article by H. C. Carver, *Annals of Mathematical Statistics*, vol. 7, p. 154.

be noticed that as we state them here they should be applied only where the class interval is unity, *i.e.*, in the u unit.

$$\text{Corrected } \mu_{2:u} = \text{uncorrected } \mu_{2:u} - \frac{1}{12}$$

$$\text{Corrected } \mu_{3:u} = \text{uncorrected } \mu_{3:u}$$

$$\text{Corrected } \mu_{4:u} = \text{uncorrected } \mu_{4:u} - \frac{1}{2} (\text{uncorrected } \mu_{2:u}) + \frac{7}{240}$$

$$\left(\frac{1}{12} = 0.08333, \frac{7}{240} = 0.02917 \right).$$

Example. For Table 18 we have

$$\text{Corrected } \mu_{2:u} = 1.720 - 0.083 = 1.637$$

$$\sigma_u = \sqrt{1.637} = 1.28$$

$$\text{Corrected } \sigma_x = 10(1.28) = 12.8\%$$

$$\begin{aligned} \text{Corrected } \mu_{4:u} &= 9.4096 - (1.72)/2 + 7/240 \\ &= 8.5788 \end{aligned}$$

$$\therefore \alpha_4 = 8.5788/(1.637)^2 = 3.20$$

The values of \bar{x} and μ_3 remain unchanged.

Sheppard's corrections are valid only for the bell-shaped types of distributions. They are not applicable to the J-shaped or U-shaped types. Moreover, they constitute a refinement which may not always be consistent with the degree of accuracy in the original data. The errors of grouping (not mistakes) are usually small compared with the errors existing in the raw data. So, it seems that little would be gained by their use in a first course. We will occasionally use them in an illustration.

CHAPTER V

MEASURES OF DISPERSION

1. Introduction. The concept of variability is fundamental today not only in the social sciences but also in the so-called exact physical sciences. Modern scientific method recognizes the existence of physical, moral, and mental inequalities. The principle of variability has come to be accepted as the natural order in social, economic, and physical phenomena. This principle is the very essence of the statistical nature of mass phenomena. In this connection, R. A. Fisher says:¹

The conception of statistics as the study of variation is the natural outcome of viewing the subject as the study of populations; for a population of individuals in all respects identical is completely described by a description of any one individual, together with the number in the group. The populations which are the object of statistical study always display variation in one or more respects. To speak of statistics as the study of variation also serves to emphasize the contrast between the aims of modern statisticians and those of their predecessors. For, until comparatively recent times, the vast majority of workers in this field appear to have had no other aim than to ascertain aggregate, or average, values. The variation itself was not an object of study, but was recognized rather as a troublesome circumstance which detracted from the value of the average. . . . Yet, from the modern point of view, the study of the causes of variation of any variable phenomena, from the yield of wheat to the intellect of man, should be begun by the examination of the variation which presents itself. The study of variation leads immediately to the concept of a frequency distribution.

It is clearly important, therefore, in studying a distribution, to describe how the variates are clustered or scattered around an average. Figure 13 shows how two distributions may even have the same mean and total frequency, yet differ considerably in variation from the mean. Such variation is commonly called dispersion, variability, or spread.

We will consider three measures of dispersion: *Quartile Deviation*,

¹ R. A. Fisher, *Statistical Methods for Research Workers*, p. 3. Oliver and Boyd, London.

Mean Deviation, and *Standard Deviation*, of which the last is by far the most important.

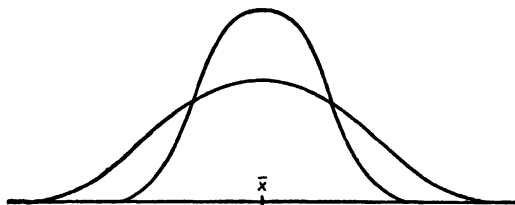


FIG. 13. TWO DISTRIBUTIONS DIFFERING IN DISPERSION

2. The Quartile Deviation. Just as the median selects one point of division, we may now take two additional points such that they, together with the median, divide the whole distribution into four equal parts. These points are called the quartile values.

The first quartile, denoted by Q_1 , is that value of x for which $\text{cum } f = N/4$. That is, one-fourth of all the variates in the distribution are smaller in value than Q_1 and three-fourths of them are larger than Q_1 . The second quartile Q_2 is that value of x for which $\text{cum } f$ is $N/2$ and is therefore the median. The third quartile, denoted by Q_3 , is that value of x for which $\text{cum } f = 3N/4$. Hence fifty per cent of the total frequency is included between Q_1 and Q_3 .

Half of the distance between Q_3 and Q_1 is called the *semi-inter-quartile range* or *quartile deviation* and will be denoted by Q . Thus,

$$(1) \quad Q = \frac{Q_3 - Q_1}{2}$$

It should be noted that the median does not necessarily come at the mid-point of $2Q$, i.e., that a distance Q laid off on either side of Q_2 would not necessarily reach to Q_1 and Q_3 . (See Figure 14.) (For a symmetrical distribution, to be considered later, this would be true.)

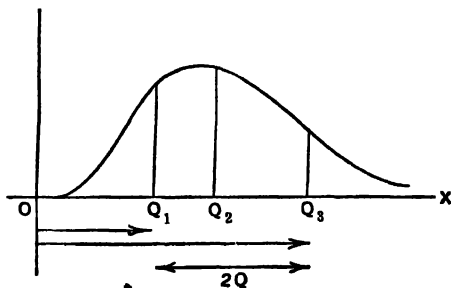


FIG. 14

As a measure of dispersion, Q gives a fairly good idea of the spread of the variates, and is suitable as such a measure in those cases where the median would be used as an average. The quartile values Q_1

and Q_3 are found, like the median, by interpolation in the cumulative frequency table.

Example. (a) Find the median and the quartile deviation for the distribution of IQ's in Table 6 (§10, Chapter I). (b) Illustrate the measures found in (a) by means of a *cum f* graph.

<i>End-x</i>	<i>Cum f</i>
54.5	0
64.5	3
74.5	24
84.5	102
← Q_1	
94.5	284
← Med.	
104.5	589
← Q_3	
114.5	798
124.5	879
134.5	900
144.5	$N = 905$

Solution:

$$N/4 = 226.25, \quad N/2 = 452.5, \quad 3N/4 = 678.75$$

$$\frac{Q_1 - 84.5}{10} = \frac{226.25 - 102}{284 - 102}, \quad Q_1 = 91.3$$

$$\frac{Q_2 - 94.5}{10} = \frac{452.5 - 284}{589 - 284}, \quad Q_2 = 100.02$$

$$\frac{Q_3 - 104.5}{10} = \frac{678.75 - 589}{798 - 589}, \quad Q_3 = 108.8$$

$$Q = \frac{Q_3 - Q_1}{2} = 8.75.$$

Figure 15 explains graphically the measures obtained by interpolation from a *cum f* table. For convenience in drawing the figure, the quartile labels are put on vertical lines. But one should remember that the quartiles are values of x and that it is the horizontal distances of the lines from the y -axis that represent these measures.

Exercises

1. Criticize the following "definitions":

$$Q_1 = \frac{N}{4}, Q_2 = \frac{N}{2}, Q_3 = \frac{3N}{4}.$$

2. Find Q_1 and Q_3 from the cumulative frequency table which you made to obtain the median for the Glasgow schoolgirl distribution. (Exercise 5 on page 52.)
3. Find the quartile deviation Q from your results in Exercise 2.
4. Find Q_1 , Q_2 , Q_3 for the distribution in Table 12, and compute Q .
5. Compute the value of the semi-interquartile range for other distributions at the direction of the instructor.

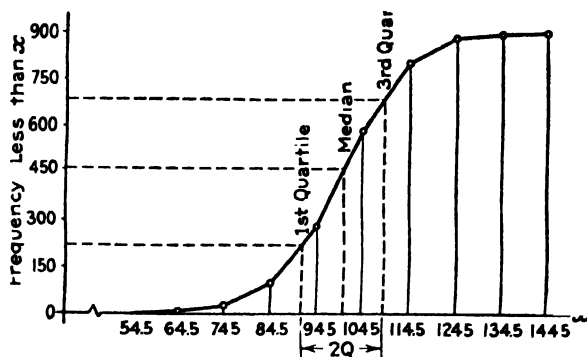


FIG. 15

6. The m th percentile P_m of a frequency distribution is that value of the variable x for which $\text{cum } f = mN/100$, where $m = 1, 2, \dots, 99$. The 10th, 20th, 30th, \dots , percentiles are called *deciles*. Therefore, the n th decile D_n is that value of x for which $\text{cum } f = nN/10$, where $n = 1, 2, \dots, 9$. Compute several percentiles and deciles of a distribution in the text.

3. Mean Deviation. As a measure of variation about a central value, it would seem appropriate to take an average of all the deviations about that central value. In the mean deviation (MD) about the mean this is precisely what we do, namely, we find the arithmetic mean of the numerical values of the deviations about the mean. In summing the deviations, their absolute values are used because regardless of whether deviations are positive or negative they have the same influence on the amount of variation. Moreover, if their algebraic signs are taken account of, the sum of such deviations is

zero (Theorem VI of Chapter III). Hence we sum them treating all deviations as positive.

In mathematical symbols, vertical bars denote absolute values, so we have¹

$$(2) \quad \mathbf{MD} = \frac{1}{N} \sum f_i |x_i - \bar{x}|,$$

if the x unit is used. When the class interval is the unit, we have

$$(3) \quad \mathbf{MD} = \frac{1}{N} \sum f_i |u_i - \bar{u}|$$

and

$$(4) \quad \mathbf{MD} (x \text{ unit}) = c \times \mathbf{MD} (u \text{ unit}).$$

It can be proved that the essentially positive function

$$y = \frac{1}{N} \sum f_i (x_i - A)^2$$

is a minimum when $A = \bar{x}$. (See Theorem II, page 99. Also by the calculus $dy/dA = 0$ when $A = \bar{x}$.) It was in a similar investigation to find the value of B for which the function

$$y = \frac{1}{N} \sum f_i |x_i - B|$$

is a minimum, that the median was discovered. When B is the median this function is a minimum.² This property of the median has some statistical importance in connection with the geographical location of centers of industry and population.³ Custom has established the use of the mean rather than the median in this measure. Hence "mean deviation" usually refers to the mean deviation from the mean. It is also called "average deviation."

¹ Since all the data are not concentrated at the midpoints of the intervals, a grouping error is involved here as in the formula for σ (§10, Chapter IV). But the mean deviation is used so infrequently that discussion here of the appropriate correction hardly seems warranted. Those who may be interested will find a more precise formula in the *Handbook of Mathematical Statistics* — Rietz and others.

² For a proof see reference 16, our Introduction.

³ See p. 85 of *Elements of Statistics* — Davis and Nelson. Principia Press.

Example. Find the mean deviation for the grades in Table 18 where the mean value of x is 72.5.

x	f	$ x - \bar{x} $	$f x - \bar{x} $
34.5	2	38	76
44.5	3	28	84
54.5	11	18	198
64.5	20	8	160
74.5	32	2	64
84.5	25	12	300
94.5	7	22	154
Total	100		1036

$$MD = \frac{1036}{100} = 10.36.$$

What was σ for this distribution?

The absolute value of a variable x' , denoted by the symbol $|x'|$, is not very tractable in mathematical operations. Therefore the mean deviation is not favored by mathematicians since it is unwieldy in the more theoretical and mathematical discussions. Its chief use is in experimental work where occasional large and erratic deviations are likely to occur. In such cases the standard deviation would tend to emphasize these deviations.

If m of the N variates are greater than the mean, \bar{x} , then the mean deviation may be written

$$\begin{aligned}
 MD &= \frac{m}{N} \left\{ (\text{sum of variates greater than } \bar{x}) - m\bar{x} \right\} \\
 &= \frac{m}{N} \left\{ \sum_{x_i > \bar{x}} f_i x_i - \bar{x} \sum_{x_i > \bar{x}} f_i \right\}
 \end{aligned}$$

The student is given a hint in Exercise 34 at the end of Part I on how to prove a similar formula for $x_i < \bar{x}$.

4. The Standard Deviation. To overcome the difficulty of negative deviations and the use of absolute value signs, the deviations about the mean may be squared and the mean of these squares taken.

To get back into the original linear units, we take the positive square root of this result, and have

$$(5) \quad \sigma_x = \left[\frac{1}{N} \sum f_i (x_i - \bar{x})^2 \right]^{1/2}$$

as defined before. The standard deviation measures the same kind of phenomenon as the mean deviation and this approach to it is frequently satisfactory to a student who otherwise finds it difficult to understand.¹

For a common type of distribution, the standard deviation is approximately twenty-five per cent greater than the mean deviation. Speaking more accurately, this is true of a normal distribution (to be considered in Chapter VI) for which the relation is $MD = \frac{4}{5}\sigma$ (approximately).

It is often convenient to have a name for "the square of the standard deviation," and for this purpose the term "variance" has been introduced. Thus σ denotes standard deviation and σ^2 denotes variance.

Although definition (5) is the basic concept which the student should have for the standard deviation, nevertheless in actual practice it is seldom desirable to compute σ directly from that definition. For a frequency distribution the method is shown in the chapter on moments. However, we will give an additional illustration here.

Example. Find the mean and the standard deviation of Table 9, using Charlier's check and Sheppard's correction.

Solution: (See Table 21, p. 88.)

$$\begin{aligned} \text{Charlier's check: } \sum f(u+1)^2 &= \sum fu^2 + 2\sum fu + N \\ 2471 &= 2365 + 2(-447) + 1000 = 2471 \end{aligned}$$

Computations:

$$\begin{aligned} \bar{x} &= 49.5 + 4(-.447) = 47.712 \text{ lbs.} \\ \mu_{2:u} &= \nu_2 - (\bar{u})^2 = 2.165 \end{aligned}$$

¹ The term "standard deviation" was proposed by Pearson and is now used by almost all English writers. As originally defined by Pearson, this is the square root of the mean of the squares of deviations *taken from the mean of the distribution*, and is not to be used when deviations are measured from any other reference point. Pearson uses the term "root-mean-square" for a similar measure when the deviations are taken around any origin other than the mean. — Walker, *History of Statistical Method*, p. 54.

Using Sheppard's corrections,

$$\text{Corrected } \mu_2 = 2.165 - .083 = 2.082$$

$$\mu_{2:z} = 16(2.082) = 33.312$$

$$\sigma_z = \sqrt{33.312} = 5.772 \text{ lbs.}$$

TABLE 21 — WEIGHTS OF GLASGOW SCHOOL CHILDREN

Weight (<i>x</i>)	<i>f</i>	<i>u</i>	<i>fu</i>	<i>fu</i> ²	<i>f(u + 1)</i> ²
29.5 lbs.	1	-5	- 5	25	16
33.5	14	-4	- 56	224	126
37.5	56	-3	-168	504	224
41.5	172	-2	-344	688	172
45.5	245	-1	-245	245	0
49.5	263	0	0	0	263
53.5	156	1	156	156	624
57.5	67	2	134	268	603
61.5	23	3	69	207	368
65.5	3	4	12	48	75
Sums	1000		-447	2365	2471
(Sums)/ <i>N</i>	1		-.447 <i>ū</i>	2.365 <i>ū</i> ²	

It will be proved later that for a certain ideal type of distribution which is often approximated in practical statistics the range $\bar{x} \pm \sigma_z$ includes about two thirds of the variates. Assuming the above distribution is of this type we could say that about two thirds of the children weighed between 42 pounds and 53.5 pounds. Such a statement assists one in comprehending certain characteristics of the data though the distribution actually may not be before him.

It is understood that the method of computation described above is to be used when the class marks are equispaced. If the class intervals are unequal we must choose $\hat{c} = 1$ unless the x 's denoting the class marks have a common factor c . When $c = 1$, u becomes $u = x - x_0$, and the work may be simplified a little by an appropriate choice of x_0 .

Exercises

1. (*Pearson*). The following data represent the percentage of ash-content in 280 wagon tests of a certain kind of coal. Find the mean and the standard deviation of the distribution:

<i>Percentage Ash-Content</i>	<i>Frequency</i>
3.0- 3.9	1
4.0- 4.9	7
5.0- 5.9	28
6.0- 6.9	78
7.0- 7.9	84
8.0- 8.9	45
9.0- 9.9	28
10.0-10.9	7
11.0-11.9	2

Ans. $\bar{x} = 7.35\%$, $\sigma_x = 1.36\%$.

2. (*Camp*). Find the mean wage and the standard deviation of the following data:

<i>Class</i>	<i>Frequency</i>
\$4.50- 5.99	43
6.00- 7.49	99
7.50- 8.99	152
9.00-10.49	178
10.50-11.99	160
12.00-13.49	40
13.50-14.99	25
15.00-16.49	3

Ans. $N = 700$, $\bar{x} = \$9.42$, $\sigma_x = \$2.19$.

3. Given $\sigma_x = 2.19$ for the following (x, f) distribution, find σ_v and σ_u for the (v, f) and (u, f) distributions, respectively.

<i>f</i>	43	99	152	178	160	40	25	3
<i>x</i>	0	1.5	3.0	4.5	6.0	7.5	9.0	10.5
<i>v</i>	0	1	2	3	4	5	6	7
<i>u</i>	-3	-2	-1	0	1	2	3	4

What relation and theorem in Chapter IV does this illustrate?

- Find the variance σ_x^2 of Table 16 (§8, Chapter III).
- Compute the value of the ratio MD/σ for the data in Exercise 1 above.
- Find the mean and standard deviation for the data in Table 10.

7. Find the mean and standard deviation for the data in Table 11.
 8. Transform the variates of the following distribution into standard units:

x	2	4	6	8	10	12	14	16	18	20
f	1	9	36	84	126	126	84	36	9	1
t	Some answers:					1/3	1	5/3	7/3	3

5. Relative Dispersions. The full significance of different values of σ can be obtained only by experience, but it is obvious that a small standard deviation indicates that the variates are closely clustered about the mean; whereas a large standard deviation indicates that these values are spread out widely from the mean. (See Figure 13.)

The size of variates usually influences not only the mean but also deviations from the mean. In other words, the magnitudes of the deviations from the mean seem to be dependent, in some degree, upon the magnitude of the mean. In comparing dispersion in distributions, we may correct for differences in the average magnitudes of positive variates by taking the ratio of the standard deviation to the mean. Thus, the quantity

$$(6) \quad V = \frac{\sigma_x}{\bar{x}}$$

is known as the *coefficient of variation*. It is obviously an abstract number, being independent of the units of measurement, and it is usually expressed as a percentage. *It is used as a measure of the uniformity of data.*

The use of (6) may be misleading in situations where the origin from which the data are measured is somewhat arbitrary. Cases in point are temperature measurements and certain psychological data. Further discussion of such limitations of (6) will be found in references 2, 14, and 15, listed in the Introduction.

6. Scaling a Distribution in Terms of σ . Suppose we lay off intervals of length σ on either side of the mean (Figure 16). Then for a certain type of distribution known as the normal curve (which will be considered in the next chapter) the following properties can be proved:

- (1) The percentage of the total frequency lying outside the range $\bar{x} \pm \sigma$ is 32% approximately.
- (2) The percentage outside $\bar{x} \pm 2\sigma$ is 5% approximately.

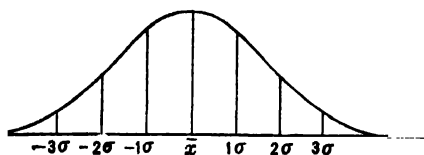


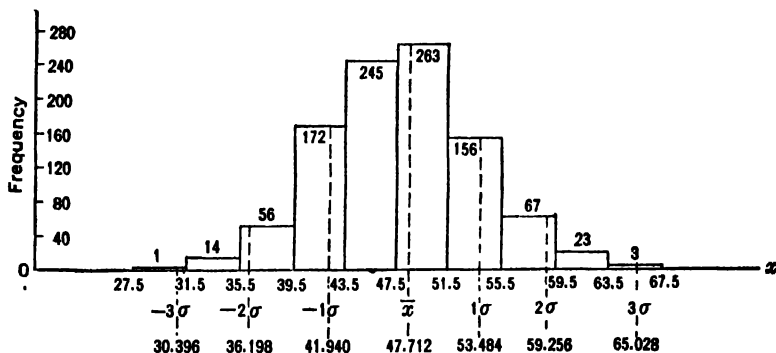
FIG. 16

(3) The range $\bar{x} \pm 3\sigma$ includes practically the whole distribution, i.e., the total range is 6σ approximately.

The student will recognize that these ranges are, in standard units, $t = \pm 1$, $t = \pm 2$, $t = \pm 3$, respectively. These results follow from the relation

$$t = \frac{x - \bar{x}}{\sigma}, \quad x = \bar{x} + t\sigma.$$

Sometimes it is important in a statistical analysis to know how nearly the given variates are distributed in accordance with the

FIG. 17 — DISTRIBUTION OF TABLE 21 SCALED OFF IN UNITS OF σ

above property of the normal curve. The distribution of Table 21 has been scaled off in this manner, with the results shown in Table 22. Figure 17 will be helpful in verifying them.

We will verify here the 34.8% given in Table 22, and the student is asked to verify the others in Exercise 2. The range $\bar{x} \pm \sigma$ (Figure 17) evidently includes all the variates represented by the two central rectangles and proportionate parts of the two adjoining rectangles. From 39.50 to 41.94 is 2.44, and since the variates are assumed to be uniformly distributed over the class interval we have $172(2.44/4) =$

104.92 for the proportionate number to be excluded in the class 39.5-43.5. Hence the number below $\bar{x} - \sigma$ is $(1 + 14 + 56 + 104.92) = 175.92$. Similarly, from 53.484 to 55.5 is 2.016, and we have $156(2.016/4) = 78.624$ as the proportionate number excluded in the class 51.5-55.5. Hence the total above $\bar{x} + \sigma$ is $(78.624 + 67 + 23 + 3) = 171.624$. So the total number outside $\bar{x} \pm \sigma$ is $(171.624 + 175.92) = 348$ or 34.8% of the 1000 variates. This re-

TABLE 22 — RESULTS OF SCALING OFF TABLE 21

$\bar{x} = 47.712$ $\sigma_x = 5.772$		Range	Frequency outside the given range	
			Number	Percent
$\bar{x} - \sigma = 41.940$	$\bar{x} + \sigma = 53.484$	$\bar{x} \pm \sigma$	348	34.8
$\bar{x} - 2\sigma = 36.198$	$\bar{x} + 2\sigma = 59.256$	$\bar{x} \pm 2\sigma$	60	6.0
$\bar{x} - 3\sigma = 30.396$	$\bar{x} + 3\sigma = 65.028$	$\bar{x} \pm 3\sigma$	3	0.3

sult could also be obtained as follows: By forming a *cum f* table and interpolating in the *end x* column we find

<i>cum f</i> at $x = 53.484$:	828
<i>cum f</i> at $x = 41.940$:	176
Number in the $(\bar{x} \pm \sigma_x)$ interval:	652
Number outside this interval:	348

7. Semi-interquartile Range in Terms of σ . The range $(Q_3 - Q_1)/2$ when expressed in units of σ has a significance in a normal distribution, as will be shown later. We will denote this by s ; hence

$$s = \frac{Q_3 - Q_1}{2\sigma}, \text{ and } s = \frac{Q}{\sigma}.$$

For the present we merely calculate its value in the exercises below.

Exercises

- Find the mean and standard deviation for the distribution of Lengths of Telephone Calls, given in Table 8 (Chapter I). Use Charlier's check.
- In the three distributions named, show that the percentages outside $\bar{x} + t\sigma$ for $t = \pm 1, \pm 2$, and ± 3 , are as stated in Table 23. Verify also the values of s .

TABLE 23

Distribution	<i>N</i>	Percent Outside			<i>s</i>
		$\bar{x} \pm \sigma$	$\bar{x} \pm 2\sigma$	$\bar{x} \pm 3\sigma$	
<i>Glasgow girls</i>	1000	34.8	6.0	0.3	0.675
<i>Telephone calls</i>	995	32.7	5.0	0.4	0.69
<i>Span</i>	2000	31.8	4.2	0.5	0.665

8. *N* Small. Ungrouped Data. When *N* is small it is seldom desirable to attempt an arrangement of the variates into a frequency distribution. Moreover, in this case, the values of α_3 and α_4 are not usually needed because the applications of these measures relate to characteristics of large distributions. Therefore, only the mean and standard deviation are usually required for a small set of ungrouped

TABLE 24 — AVERAGE YIELDS OF CORN IN BUSHELS PER ACRE
FOR A CERTAIN SECTION IN ILLINOIS FROM 1901-1920

Year	Yield (<i>x</i>)	<i>n</i>	<i>n</i> ²
1901	21	-15	225
1902	39	3	9
1903	32	-4	16
1904	37	1	1
1905	40	4	16
1906	36	0	0
1907	36	0	0
1908	32	-4	16
1909	36	0	0
1910	39	3	9
1911	33	-3	9
1912	40	4	16
1913	27	-9	81
1914	29	-7	49
1915	36	0	0
1916	30	-6	36
1917	38	2	4
1918	36	0	0
1919	36	0	0
1920	35	-1	1
Totals	<i>N</i> = 20	-32	488

data. The following methods will help the student become familiar with the several formulas for σ , which may be used in this case.

Method I. The indirect method involving the u unit may still be used for finding the first and second moments. Since each variate is being treated separately $f = 1$, and we compute the values of

$\nu_r = \frac{1}{N} \sum u^r$ for $r = 1$ and 2. If the values of x are unequally spaced

we take $c = 1$ and let $u = x - x_0$ which changes the origin but not the units. In other words, the procedure is the same as for a frequency distribution except that $f = 1$ and $c = 1$.

Example. Find the mean and standard deviation for Table 24. $N = 20$. We choose $x_0 = 36$.

TABLE 25

x	$x' = x - \bar{x}$	x'^2
21	-13.4	179.56
27	- 7.4	54.76
29	- 5.4	29.16
30	- 4.4	19.36
32	- 2.4	5.76
32	- 2.4	5.76
33	- 1.4	1.96
35	0.6	.36
36	1.6	2.56
36	1.6	2.56
36	1.6	2.56
36	1.6	2.56
37	2.6	6.76
38	3.6	12.96
39	4.6	21.16
39	4.6	21.16
40	5.6	31.36
40	5.6	31.36
688	$\sum x' = 73.6$	436.80

Computations:

$$\nu_1 = \bar{u} = -\frac{32}{20} = -1.6; \quad \bar{x} = x_0 + \bar{u} = 36 - 1.6 \\ = 34.4 \text{ bushels}$$

$$\nu_2 = \frac{488}{20} = 24.40; \quad \mu_2 = \nu_2 - \bar{u}^2 = 21.84.$$

Therefore,

$$\sigma_x = \sigma_u = \sqrt{21.84} = 4.67 \text{ bushels.}$$

Method II. When $f = 1$, formula (5) becomes

$$(7) \quad \sigma_x = \left[\frac{1}{N} \sum (x_i - \bar{x})^2 \right]^{1/2},$$

and sometimes it is best to compute the standard deviation directly from this definition, without the use of the u unit. Thus the origin is placed at the mean and all indirect methods are abandoned. If the mean deviation is also desired, clearly this method should be used. It is exemplified in Table 25 for the preceding example, and the variates have been arranged in order of magnitude.

$$\bar{x} = \frac{688}{20} = 34.4 \text{ bushels}$$

$$\sigma_x^2 = \frac{436.80}{20} = 21.84$$

$$\sigma_x = 4.67 \text{ bushels}$$

$$\text{MD} = \frac{1}{N} \sum |x'| = \frac{73.6}{20} = 3.68 \text{ bushels.}$$

Method III. From the relation

$$\mu_2 = \nu_2 - (\nu_1)^2$$

we have

$$\mu_2 = \frac{1}{N} \sum x^2 - \bar{x}^2$$

when $f = 1$. Therefore σ may be written

$$(8) \quad \sigma_x = \left[\frac{1}{N} \sum x^2 - \bar{x}^2 \right]^{1/2}.$$

TABLE 26

x	x^2
21	441
27	729
29	841
30	900
32	1024
32	1024
33	1089
35	1225
36	1296
36	1296
36	1296
36	1296
36	1296
37	1369
38	1444
39	1521
39	1521
40	1600
40	1600
688	24,104

This method is perhaps the best when the values of x are not large or when a table of squares is available. It is illustrated below for the preceding example. (See Table 26.)

Computations:

$$\bar{x} = \frac{1}{N} \sum x = \frac{688}{20} = 34.4 \text{ bushels}$$

$$\bar{x}^2 = (34.4)^2 = 1183.36$$

$$\frac{1}{N} \sum x^2 = \frac{24104}{20} = 1205.20$$

$$\begin{aligned} \sigma_x &= [1205.20 - 1183.36]^{1/2} \\ &= (21.84)^{1/2} \\ &= 4.67 \text{ bushels.} \end{aligned}$$

Miscellaneous Exercises

- (a) Verify that the algebraic sum of the numbers in the x' column of Table 25 is zero.
- (b) Verify the value of mean deviation given for Table 25.

2. Using your own judgment as to the most appropriate method, find the mean and standard deviation for each of the two sets of data, x_1 and x_2 :

														Answers
x_1	88	95	68	73	75	88	57	68	62	79	73	74	78	$\bar{x}_1 = 69.80$ $\sigma_1 = 12.13$
	80	57	65	69	74	78	72	59	47	56	67	43		
x_2	82	86	75	78	72	79	63	65	67	75	68	70	79	$\bar{x}_2 = 67.64$ $\sigma_2 = 12.68$
	78	51	58	65	69	68	83	80	42	43	48	47		

3. Complete the computations and find the mean and variance of the following distribution:

Data		Computations		
y	f	v	vf	v^2f
67.86	7	-16.94		
72.14	14	-12.66		
81.87	32	- 2.93		
84.80 ← y_0	49	0		
85.73	55	.93		
90.92	54	6.12		
95.57	35	10.77		
105.00	14	20.20		

Hint. Here we let $v = y - y_0$. Then $\bar{y} = \bar{v} + y_0$, and $\sigma_y^2 = \sigma_v^2$ since $c = 1$. (See Theorem on p. 69.)

Ans. $\bar{y} = 87.31$, $\sigma_y^2 = 56.66$.

4. Data have been gathered showing the points scored on a mental test by 290 prospective employees and the per cent of standard production attained by these same 290 persons after being employed.¹ The following statistics were obtained:

Mental test: mean = 43.33 pts.

$\sigma = 9.25$ pts.

Productive ability: mean = 92.02%

$\sigma = 24.47\%$

- (a) Compare the relative dispersion in mental test and productive ability.
 (b) What factors, other than mental level, may have affected dispersion under factory conditions?

¹ Wembridge, "Experiment and Statistics in the Selection of Employees," *Journal of the American Statistical Association*, March 1923, p. 605.

5. Read and abstract the article "Variability," *Journal of Educational Research*, vol. 4, no. 3, pp. 221-26.
6. Find the median for Table 26.
7. Find \bar{x} , σ_x , MD, and Q for the following distribution.

<i>mid-x</i>	2	4	6	8	10
<i>f</i>	1	4	6	4	1

8. Show that (8) may be written as follows:

$$\sigma_x = \frac{[N\sum x^2 - (\sum x)^2]^{1/2}}{N}.$$

9. If the variates are all equal, say each $x_i = k$, show that $\bar{x} = k$ and $\sigma = 0$.
10. For a set of ungrouped data it is found that $N = 15$, $\sum x = 480$, $\sum x^2 = 15,735$. Find \bar{x} and σ_x .
11. Find the variance of the following data.

5.7 6.2 6.5 6.0 6.3 5.8 5.7 6.0 6.0 5.8

Ans. $\sigma_x^2 = .064$.

12. Prove the identity:

$$\begin{aligned} (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2 \\ = (x_1^2 + x_2^2 + \cdots + x_N^2) - N\bar{x}^2. \end{aligned}$$

13. Compute the mean deviation (from the mean) for the following data:

<i>x</i>	2	4	6	8	10	17
<i>f</i>	1	6	10	7	2	2

Ans. MD = 33/14.

14. Verify the identity (where \bar{x} is the mean of x_1 and x_2):

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 = \frac{1}{2}(x_1 - x_2)^2,$$

and thus show that, for two variates,

$$|x_1 - x_2|$$

15. Verify the identity (where \bar{x} is the mean of x_1, x_2, x_3):

$$3(x_1 - x_2)^2 + (x_1 + x_2 - 2x_3)^2 = 6 \sum_{i=1}^3 (x_i - \bar{x})^2.$$

9. The Standard Deviation of the Combination of Sets. The following theorems involving σ are interesting in themselves and have useful applications.

The relation $\mu_2 = \nu_2 - \nu_1^2$ is true in a more general sense than we have previously used. Its generalized meaning will be revealed in our first theorem.

Theorem I. *The second moment about the mean equals the second moment about an arbitrary point $P(x_0, 0)$ minus the square of the distance between the mean and P .*

Stated in symbols the theorem may be clearer. Suppose we have a set of N variates whose mean is \bar{x} . Graphically, \bar{x} is a point on the x -axis. Then if P is any other point on the x -axis, according to Theorem I we have

$$(9) \quad \frac{1}{N} \sum (x - \bar{x})^2 = \frac{1}{N} \sum (x - x_0)^2 - (\bar{x} - x_0)^2.$$

To prove this relation we may write

$$(x - \bar{x}) = (x - x_0) - (\bar{x} - x_0).$$

Then

$$\frac{1}{N} \sum (x - \bar{x})^2 = \frac{1}{N} \sum [(x - x_0) - (\bar{x} - x_0)]^2,$$

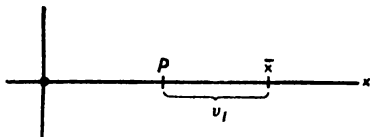
the right member of which simplifies into the right member of (9).

The generality of the theorem consists in extending the original definition of ν_2 and ν_1 so that they refer to moments about any point P on the x -axis (except \bar{x}), and not merely about zero. Thus now,

$\nu_2 = \frac{1}{N} \sum (x - x_0)^2$. If we take $x_0 = 0$ we have the original defini-

tion of ν_2 . Also, when P moves back to zero, we see that ν_1 becomes \bar{x} . In other words, the original definitions of the ν 's are merely the more general definitions when

zero is the value chosen for the arbitrary point. (See (1a) of Chapter IV.)



Theorem II. *The sum of the squares of deviations of the variates from their mean is less than the sum of the squares of the deviations of the variates from any other value. Therefore σ is less than any similar "root-mean-square."*

The proof consists in showing that $\mu_2 < \nu_2$ which is left to the student as an exercise.

Theorem III. *Let there be one set of n_1 variates x_{1i} ($i = 1, 2, \dots, n_1$) and another set of n_2 variates x_{2i} ($i = 1, 2, \dots, n_2$) and let \bar{x} be the mean of the combined sets (Theorem VIII, Chapter III). The variance σ^2 of the set formed by the combination of these two sets is given by the following formula:*

$$(10) \quad N\sigma^2 = \sum_1^{n_1} (x_{1i} - \bar{x})^2 + \sum_1^{n_2} (x_{2i} - \bar{x})^2$$

where

$$N = n_1 + n_2.$$

Proof: The proof consists in showing that

$$\sum_1^{n_1} (x_{1i} - \bar{x})^2 + \sum_1^{n_2} (x_{2i} - \bar{x})^2 = \sum_1^{n_1+n_2} (x_i - \bar{x})^2$$

which is left as an exercise for the student.

The above theorem is not very important in itself but it is useful in proving the next theorem which gives the relation between the variance of a composite set and the variances of sub-sets.

Theorem IV. *Let the frequency, mean, and standard deviation be denoted by n_1 , \bar{x}_1 , and σ_1 for one set of variates and by n_2 , \bar{x}_2 , and σ_2 for a second set. The variance σ^2 of the composite set is given by the following relation:*

$$N\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2$$

where $N = n_1 + n_2$, $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$, and \bar{x} is the mean of the composite set.

Proof: For the n_1 set, \bar{x} may be regarded as an arbitrary point P . Hence by Theorem I we have

$$\frac{1}{n_1} \sum_1^{n_1} (x_{1i} - \bar{x}_1)^2 = \frac{1}{n_1} \sum_1^{n_1} (x_{1i} - \bar{x})^2 - (\bar{x}_1 - \bar{x})^2.$$

Multiplying through by n_1 this becomes

$$(11) \quad n_1\sigma_1^2 = \sum_1^{n_1} (x_{1i} - \bar{x})^2 - n_1d_1^2.$$

Similarly for the n_2 group we have

$$(12) \quad n_2\sigma_2^2 = \sum_1^{n_2} (x_{2i} - \bar{x})^2 - n_2d_2^2.$$

Adding (11) and (12), and using (10), we obtain

$$n_1\sigma_1^2 + n_2\sigma_2^2 = N\sigma^2 - n_1d_1^2 - n_2d_2^2.$$

Hence,

$$(13) \quad N\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2.$$

For k sets combined into a single set we can generalize (13) into the following relation:

$$(14) \quad N\sigma^2 = \sum_1^k n_i\sigma_i^2 + \sum_1^k n_id_i^2$$

where $N = \sum_1^k n_i$, and $d_i = \bar{x}_i - \bar{x}$. It is interesting to observe that $\frac{1}{N} \sum_1^k n_id_i^2$ is the variance of the means of the sub-sets. Thus we have the important relation

$$(14a) \quad \sigma^2 = \frac{1}{N} \sum_1^k n_i\sigma_i^2 + \sigma_{\bar{x}_i}^2$$

which shows that the total variance may be broken up into two parts, one of which is the weighted mean of the variances in the sub-sets and the other is the variance of their means. These two parts are sometimes called the average variance *within* classes and the variance *between* the means of the classes. They become very important in the "Analysis of Variance" (which is explained in Part II).

COROLLARY I. *Equation (13) may be written in the following form:*

$$(15) \quad N\sigma^2 = n_1(\sigma_1^2 + \bar{x}_1^2) + n_2(\sigma_2^2 + \bar{x}_2^2) - N\bar{x}^2.$$

Proof: Since

$$n_1d_1^2 = n_1(\bar{x}_1 - \bar{x})^2 = n_1\bar{x}_1^2 - (2n_1\bar{x}_1\bar{x} - n_1\bar{x}^2)$$

and

$$n_2d_2^2 = n_2(\bar{x}_2 - \bar{x})^2 = n_2\bar{x}_2^2 - (2n_2\bar{x}_2\bar{x} - n_2\bar{x}^2)$$

the proof consists in showing that the sum of the terms in the end parentheses above reduces to $N\bar{x}^2$. Rearranging these terms their sum is

$$2\bar{x}(n_1\bar{x}_1 + n_2\bar{x}_2) - \bar{x}^2(n_1 + n_2),$$

which by Theorem VIII (Chapter III) becomes

$$2\bar{x}N\bar{x} - \bar{x}^2N = N\bar{x}^2.$$

Generalizing for k groups, (15) becomes

$$(16) \quad N\sigma^2 = \sum_1^k n_i(\sigma_i^2 + \bar{x}_i^2) - N\bar{x}^2.$$

COROLLARY II. *Equation (13) may also be written in the form:*

$$(17) \quad N\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + \frac{n_1n_2}{N}(\bar{x}_1 - \bar{x}_2)^2.$$

The proof consists in showing that

$$n_1d_1^2 + n_2d_2^2 = \frac{n_1n_2}{N}(\bar{x}_1 - \bar{x}_2)^2.$$

This is left as an exercise.

For purposes of computation, (17) may be more convenient than either (13) or (15) because it does not require \bar{x} , but it does not lend itself to a generalization for k sets. Generalizations may be useful both for computing and for theoretical purposes. Formula (14) is particularly useful in developing the theory of a later section.

For convenience, the formulas of Theorem VIII, Chapter III, are repeated here:

$$(18) \quad \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2},$$

$$(18a) \quad \bar{x} = \frac{1}{N} \sum_1^k n_i\bar{x}_i, \quad N = \sum_1^k n_i.$$

Theorem V. *Consider k sets. Suppose the second moment of each set is taken about the mean, \bar{x} , of the combined sets. Let $\nu_2^{(i)}$ represent this moment for the i th set. Then the variance σ^2 for the combined sets is given by*

$$(19) \quad N\sigma^2 = n_1\nu_2^{(1)} + n_2\nu_2^{(2)} + \cdots + n_k\nu_2^{(k)} = \sum_1^k n_i\nu_2^{(i)}$$

when n_i represents the frequency in the i th set and $\sum_1^k n_i = N$.

Proof: We may write (10) in the form

$$N\sigma^2 = n_1\nu_2^{(1)} + n_2\nu_2^{(2)}.$$

So, generalizing this form of (10) for k sets, we obtain (19).

The next theorem gives the standard deviation of the distribution formed by the first N integers, that is, when $x = 1, 2, 3 \cdots, N$. It is

useful in cases when the variates are recorded not by measurements but by their respective positions when ranked in order with respect to some character or property.

Theorem VI. *The standard deviation σ of the first N natural numbers is given by*

$$(20) \quad \sigma = \left[\frac{(N^2 - 1)}{12} \right]^{1/2}.$$

Proof: By a fundamental definition we have

$$\sigma^2 = \frac{1}{N} \sum_{x=1}^N x^2 - \left[\frac{1}{N} \sum_{x=1}^N x \right]^2$$

and by Theorems IV and V of Chapter III, this becomes

$$\sigma^2 = \frac{1}{6}(N+1)(2N+1) - \frac{1}{4}(N+1)^2$$

which reduces to

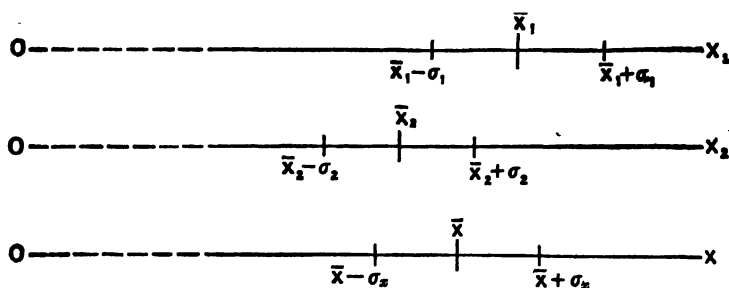
$$\sigma^2 = \frac{N^2 - 1}{12},$$

whence we obtain (20).

10. Graphical Representation. We have shown that, if certain statistics are given for two sub-sets,

<i>Sub-sets</i>	n_1	\bar{x}_1	σ_1
	n_2	\bar{x}_2	σ_2
<i>Composite set</i>	N	\bar{x}	σ_x

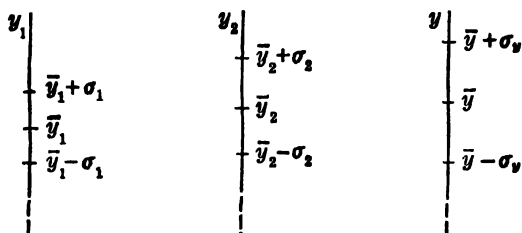
the corresponding statistics for the composite set may be obtained by means of (13) and (18a). We have been thinking of these statistics as relating to distributions in the x -direction. The following diagrams show how the means and standard deviations of three such distributions may be represented geometrically by the points whose ordinates are zero and whose abscissas are, respectively, \bar{x}_1 , $(\bar{x}_1 \pm \sigma_1)$; \bar{x}_2 , $(\bar{x}_2 \pm \sigma_2)$; and \bar{x} , $(\bar{x} \pm \sigma_x)$. The points are plotted on three different axes to avoid confusion, but they are to be thought of as being referred to the same origin and plotted on the same scale.



It should be clear that Theorems I-IV (§9) will apply to distributions in the y -direction as well as in the x -direction. In particular, it is obvious that (13) and (18a) hold if we replace x by y . Then the graphical representation of the means and standard deviations

Sub-sets		Composite set
n_1	n_2	N
\bar{y}_1	\bar{y}_2	\bar{y}
σ_1	σ_2	σ_y

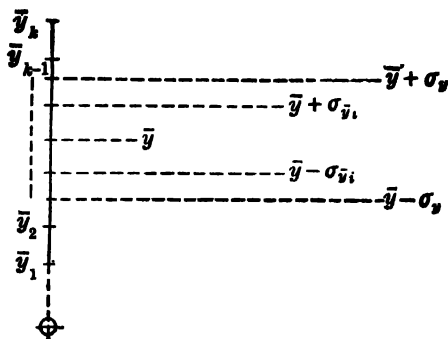
is shown below.



It will be helpful to discuss one more notion in this connection. Suppose the y composite set is made up of k sub-sets and the means $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_k$, of these sub-sets are plotted on the y -axis as shown by the labels on the left side of the axis in the figure on page 105.

We will denote the standard deviation of these means by $\sigma_{\bar{y}_i}$. Then the points \bar{y} , $(\bar{y} \pm \sigma_{\bar{y}_i})$, and $(\bar{y} \pm \sigma_y)$, may be plotted as shown. We would expect less variability among the means of the sub-sets than among the y 's of the composite set, that is, that $\sigma_{\bar{y}_i}$ would be

less than σ_y . It is clear that (14) and (14a) hold when x is replaced by y .



A grasp of these notions will help in the analysis of Table 27 which the student is asked to make in problems 5 and 6 below.

Exercises

- (a) Show that $v_1 = \frac{1}{N} \sum (x_i - x_0) = (\bar{x} - x_0)$.
 (b) Derive equations (9) and (13). If $n_1 = n_2$, what does (13) reduce to?
- Given the following information about two sets of data:

I	II
$n_1 = 20$	$n_2 = 30$
$\bar{x}_1 = 25$	$\bar{x}_2 = 20$
$\sigma_1^2 = 5$	$\sigma_2^2 = 4$

Find the mean and variance of the composite set.

- Think of the two groups in Exercise 2, page 97, as combined into a single set.
 (a) Find the mean of the combined set by formula (18).
 (b) Find the standard deviation of the combined set using result of (a) and formula (13). *Ans.* $\bar{x} = 68.72$, $\sigma = 12.45$.
- Using Theorem VI find the mean and standard deviation of the first 25 natural numbers.
- Consider Table 27. Observe that the first and last columns form a frequency distribution and that columns (1) to (8) are subdistributions whose totals add up to $N = 260$ which is also the sum of the last column. Let n_i represent the frequency in the i th column and answer the following questions: $n_1 = ?$, $n_4 = ?$, $n_8 = ?$, $\sum_{i=1}^8 n_i = ?$ Let \bar{y}_i and σ_i^2 represent mean and variance in the i th column. Find the mean and variance of each of the columns (1) to (8), first in v units where $v = (y - 85)/10$. Check your answers with those given at the bottom of the table.

TABLE 27

y	v	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	f
125	4					2	3	2		7
115	3			1	3	1	4	4	4	17
105	2			5	7	8	11	8	7	46
95	1		2	1	10	12	9	8	2	44
85	0	1	3	12	11	7	12	7	1	54
75	-1	2	1	5	6	16	8	5		43
65	-2	2	5	5	8	8	6	1		35
55	-3	2	3	3	4	1	1			14
	n_i	7	14	32	49	55	54	35	14	$N = 260$
Answers	\bar{y}_i	67.86	72.14	81.87	84.80	85.73	90.92	95.57	105.0	$\bar{y} = 87.31$
	σ_i^2	106.12	191.83	246.48	283.63	257.65	294.51	222.53	71.43	$\sigma_y^2 = 303.11$

6. Using formulas (18a) and (14) find the mean \bar{y} and variance σ_y^2 of the total distribution in Table 27 and check your answers with those given at the bottom of the last column.

Hint. The student will observe that the means, \bar{y}_i , of the columns in Table 27 are the values denoted by y in Exercise 3, page 97. The weighted mean of these mean values is the mean of the whole table. That is, from (18a),

$$\bar{y} = \frac{1}{N} \sum_1^k n_i \bar{y}_i \\ = 87.31.$$

The answer 56.66 (Exercise 3) is the variance, $\sigma_{\bar{y}_i}^2$, of the *means* of the columns of Table 27 and is not to be confused with the variance σ_y^2 of the whole table. In using (14), σ^2 is the variance of the whole table, σ_i^2 is the variance of the i th column, and the expression $\sum_1^k n_i d_i^2$ equals $N \sigma_{\bar{y}_i}^2$ where $\sigma_{\bar{y}_i}^2$ is the variance of the means of the columns since now $d_i = \bar{y}_i - \bar{y}$.

7. In Theorem V (§9) show that

$$\nu_2^{(i)} = \sigma_i^2 + d_i^2.$$

Hence prove that (19) may be derived from (14) by showing that (14) may be written as follows:

$$N\sigma^2 = \sum_1^k n_i (\sigma_i^2 + d_i^2).$$

8. (a) Derive the following relation from (18a),

$$\bar{x}_1 = \frac{1}{n_1} \left[N\bar{x} - \sum_{i=2}^k n_i \bar{x}_i \right].$$

What does this formula become when $k = 2$?

- (b) Derive the following relation from (15),

$$\sigma_1^2 = \frac{1}{n_1} \left[N(\sigma^2 + \bar{x}^2) - n_2(\sigma_2^2 + \bar{x}_2^2) \right] - \bar{x}_1^2.$$

9. In a certain distribution of $N = 25$ measurements it was found that $\bar{x} = 56$ inches and $\sigma = 2$ inches. After these results were computed it was discovered that a mistake had been made in one of the measurements which was recorded as 64 inches. Find the mean and standard deviation if the incorrect variate, 64, is omitted.

Hint. Let $n_1 = 24$, $n_2 = 1$. Then $\bar{x}_2 = 64$ and $\sigma_2 = 0$. To find \bar{x}_1 and σ_1 use formulas in Exercise 8 above.

10. If two or more variates are deleted from a distribution for which N , \bar{x} , and σ are given, show how to compute the mean and variance of the remaining variates.

11. Consider a composite set consisting of k sub-sets and let σ_i^2 and n_i denote, respectively, the variance and number of variates in the i th sub-set, and $N = \sum_1^k n_i$.

(a) If the sub-sets have equal means, show that the variance of the composite set is given by

$$\sigma^2 = \frac{1}{N} \sum_1^k n_i \sigma_i^2.$$

(b) If the sub-sets each contain the same number of variates and have equal means, show that

$$\sigma^2 = \frac{1}{k} \sum_1^k \sigma_i^2.$$

CHAPTER VI

TYPES OF DISTRIBUTIONS. THE NORMAL CURVE

1. Skewness and Kurtosis. The shapes of frequency distributions are not all alike. Unimodal distributions may differ in two ways with respect to form. These differences can be described more easily if we think in terms of frequency curves. The curve may be quite

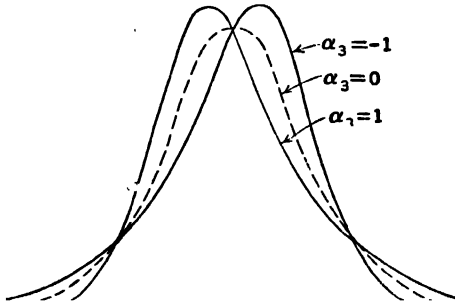


FIG. 18

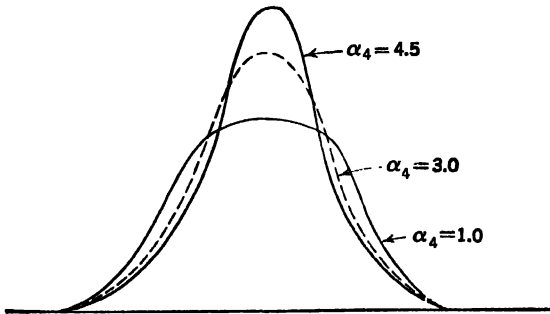


FIG. 19

symmetrical, or it may be skew, bulging out on one side more than on the other. Secondly, the top of the curve may be narrow and peaked, or it may be somewhat flat giving a mound-shape effect.

The mean and standard deviation are not sufficient to detect these characteristics, so we need other measures to describe them. Con-

TABLE 28

	A	B	C
u	f	f	f
-3	0	1	0
-2	3	1	1
-1	6	5	10
0	7	11	6
1	6	5	5
2	3	1	2
3	0	1	1
Sums	25	25	25

sider, for example, the three distributions of the weights (in class units)¹ of different breeds of mice 120-130 days old given in Table 28. Experiments on mice are important in cancer research. These distributions are, however, somewhat fictitious, being adapted from some actual data for purposes of illustration.

The student may easily verify that for each of these distributions we find the same mean and standard deviation, namely,

$$\bar{u} = 0, \quad \sigma_u = 1.2.$$

One may see from their histograms that these distributions are essentially different in shape even though they all have the same mean and standard deviation. These differences would be more pronounced if N were so large that the shapes approached a regular and smooth form. Such a large value is called the "population" or "universe" and the value of N that we usually have at hand is a "sample."

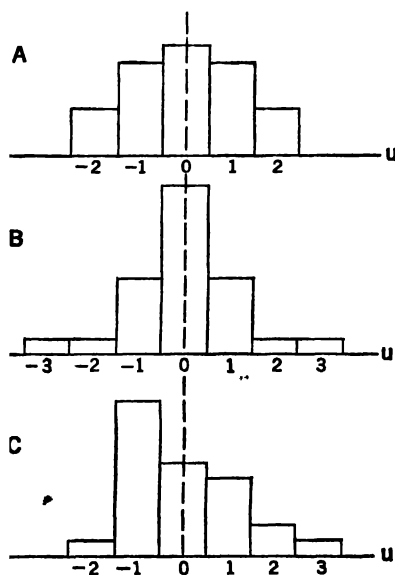


FIG. 20

¹ Neither the original units nor the class interval need concern us here.

Lack of symmetry in a distribution is known as "skewness." This characteristic is measured by α_3 . If a distribution is symmetrical $\alpha_3 = 0$, but α_3 may be positive or negative depending upon whether the long tail of the distribution extends to the right or the left of the mean. (See Figure 18.)

Figure 19 exhibits curves with different degrees of flatness or peakedness. The flatness that we are now describing is in the neighborhood of the mode and is not to be confused with the flatness of a curve as a whole which is due to spread or dispersion. The curves in Figure 19 all have the same spread. So their flatness depends upon the relative amount of material in the vicinity of the mode. This characteristic of a curve is called "kurtosis" and is measured by α_4 . By the calculus it can be demonstrated that $\alpha_4 = 3$ for a certain type of distribution which is called the normal curve. A frequency curve is said to have positive kurtosis if $\alpha_4 > 3$ and negative kurtosis if $\alpha_4 < 3$. It seems, however, that any combination of kurtosis and peakedness may occur.¹ The values of α_3 and α_4 computed for an observed distribution are useful in selecting the curve which will best represent the type to which that distribution belongs.

Both α_3 and α_4 are abstract numbers and therefore skewness and kurtosis in different distributions may be compared by these measures. Therefore our definitions are

$$(1) \quad \begin{cases} \alpha_3 \text{ is a measure of skewness,} \\ \alpha_4 \text{ is a measure of kurtosis.} \end{cases}$$

For an unsymmetrical distribution the distance between the mean and mode may be used to measure the degree of asymmetry or skewness, because the mean and mode coincide in a symmetrical distribution. Since we wish any measure of skewness to be a pure number, we would express this distance in units of the standard deviation, thus $(\text{mean} - \text{mode})/\sigma$. Now it happens that there is a certain curve known as Pearson's Type III which is used to represent certain

¹ A Common Error Concerning Kurtosis — I. Kaplansky. *J. Amer. Stat. Assoc.*, vol. 40, p. 259, June 1945. In this connection, Professor I. W. Burr comments: "The shape of the hump of the curve has less influence on α_4 than does the length of the tails. In Figure 19, the curve with $\alpha_4 = 4.5$ should have the longest tails."

skew distributions, and it can be shown by higher mathematics that, for this curve,

$$(2) \quad \frac{\text{mean} - \text{mode}}{\sigma} = \frac{\alpha_3}{2}.$$

So this relation¹ may be used as a formula for obtaining the approximate mode.

Exercise

Find α_3 and α_4 for each of the distributions A , B , and C , in Table 28.

2. Frequency Curves. As the student extends his experience he finds several types of distributions. It is important in certain problems to differentiate between them. Differences in type lead to the study of frequency curves. There are several standard curves to represent the different types of distributions that arise in practical statistics.² Each of these is specified by a mathematical function $y = f(x)$ where $f(x)$ is a general symbol for any function of x . It is, of course, a different expression for each of the different curves.

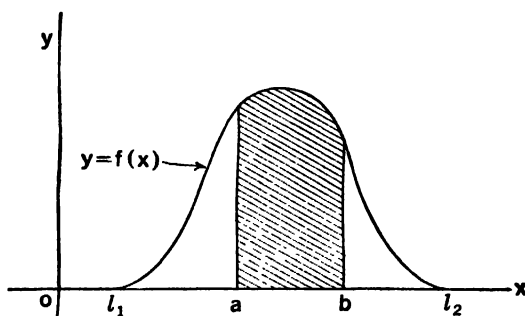


FIG. 21

Such functions are also called distribution functions. A complete discussion of this subject belongs to the field of advanced statistics. However, there are some simple concepts relating to frequency curves which will be useful in our work.

If a frequency curve is used to represent a given distribution, the total area under the curve corresponds to the total frequency N ,

¹ Because of this relation some writers use $\alpha_3/2$ as a measure of skewness instead of α_3 . Also some authors adopt a different convention as to sign, defining skewness as negative when the mean is greater than the mode.

² See Chapter III, Part II.

and therefore the partial area under the curve between the ordinates erected at $x = a$ and $x = b$ (Figure 21) represents the number of variates with measurement or character between a and b . The limits between which the theoretical distribution ranges are denoted by l_1 and l_2 . It is often convenient and causes no loss of generality to suppose that the total area under the curve is unity or 100%, in which case the partial area between a and b represents the percentage of variates having the given character.

In mathematical language the "area under $f(x)$ between a and b " is called the "integral of $f(x)$ from a to b ," and is denoted by the symbol

$$\int_a^b f(x) dx.$$

However, we will abbreviate this symbol and use merely \int_a^b to denote such an area.

Without attempting to be rigorous, we may say that the total area under the curve is the limit of the area of the appropriate histogram whose rectangles have bases Δx and altitudes $f(x)$, as Δx is taken smaller and smaller and approaches zero. Thus

$$\int f(x) dx = \lim_{\Delta x \rightarrow 0} \sum f(x) \Delta x.$$

The integral sign \int is a conventionalized S and denotes the sum of elements of area with bases dx and altitudes $y = f(x)$. The letters written at the top and bottom of \int denote the range over which the sum is to be taken. Therefore the notation $\int_a^b y dx$ or $\int_a^b f(x) dx$ represents the area which is bounded by the curve $y = f(x)$, the ordinates at $x = a$ and $x = b$, and the x -axis. (Figure 21.)

The integral of $y = f(x)$ from l_1 to l_2 denotes the total frequency N . Therefore,

$$N = \int_{l_1}^{l_2}.$$

Hence, the proportion of variates having some character x , such that $a \leq x \leq b$, is given by $\frac{1}{N} \int_a^b$. If N is taken as unity or 100%, then

\int_a^b denotes the percentage of variates having the given character.

The integral represented by this symbol also denotes the probability that a variate chosen at random from the universe $y = f(x)$ will have a value between a and b .

3. The Normal Curve. Perhaps the most important of all frequency curves is the so-called normal¹ curve whose equation may be written

$$(3) \quad y = Ke^{-h^2(x-m)^2},$$

where K , h^2 , and m represent numbers whose significance will be explained presently. The curve is bell-shaped and is symmetrical about the line $x = m$. It was first discovered by a famous French mathematician, De Moivre, over two hundred years ago and published in 1733. He obtained it while working on certain problems in games of chance which were proposed to him by the gamblers of his day. Because of this origin and because the data from certain coin- and dice-throwing experiments closely approach it in form, it is often called the normal probability curve. Actual statistical use of the normal curve began with the work of the famous mathematical astronomers, Laplace (1749-1827) and Gauss (1777-1855), each of whom derived it independently and presumably without knowing of De Moivre's treatment.² They found that it represented very well the errors of observation in the physical sciences. For this reason it has been called the normal curve of error, where error is used in the sense of a deviation from the true value. Since that time experience has shown that it serves quite well to describe many of the distributions which arise in the fields of biology, education, and sociology. Much of the theory of statistics is built around it.

The calculus is required to define the moments of a theoretical distribution specified by a frequency curve $y = f(x)$. (These definitions are given in Part II.) It turns out that *the mean of the distribution specified by (3) is m and its variance is $1/(2h^2)$* . The constant K is determined so that the area under the curve shall have some relevant value. In describing an observed distribution by

¹ The term "normal" used here should not be interpreted to mean that other types of distribution are abnormal.

² For a more extensive history see (a) "Bi-centenary of the Normal Curve," *Jour. Amer. Statistical Assoc.*, vol. 29 (1934), pp. 72-75. (b) "Mathematical Statistics" (Carus Monograph) — Rietz, Ch. 3.

means of a normal curve, we wish to have the number of area units under the curve (3) equal to the number N of observed variates. When this condition is imposed, $K = Nh/\sqrt{\pi}$ and we see that K depends also on h . If we adopt the same notation¹ here as we used for an observed distribution, we have

$$m = \bar{x}, \quad h^2 = \frac{1}{2\sigma_x^2}, \quad K = \frac{N}{\sigma_x \sqrt{2\pi}}.$$

Upon making these replacements, (3) becomes

$$(3a) \quad y = \frac{N}{\sigma_x \sqrt{2\pi}} e^{-(x-\bar{x})^2/2\sigma_x^2}.$$

4. Standard Form. The letters π and e represent numbers which always have the same values (see §1, Chapter I). But each of the letters m , h , and K may take on different values in different situations. Such constants are called parameters, and (3) really represents a *family* of curves. Similarly, in (3a), \bar{x} , σ , and N are parameters. For assigned values they determine, respectively, the position of the curve along the x -axis, its steepness, and its "size" but they do not have anything to do with its fundamental characteristics (*i.e.*, those properties which differentiate it from all other curves). In order to study these characteristic properties it is convenient to represent the curve by an equation which will be independent of the parameters; in other words, to eliminate them from the equation by a transformation. This is accomplished by considering the total area under the curve as unity, taking the origin at the mean, and using the standard deviation as the unit of horizontal measurement. In mathematical language this means that we set $N = 1$, and $t = (x - \bar{x})/\sigma_x$. We will denote the resulting function by $\phi(t)$, that is,

$$(4) \quad \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

which is called the *standard form* of the normal curve.

A variable, t , which is distributed in accord with (4) is said to be normally distributed with mean zero and unit standard deviation.

Just as coördinates of points on the curve are denoted by (x, y)

¹ In the theory of sampling, Part II, it is necessary to distinguish the moments of a sample from those of the parent universe by the use of different symbols.

in the case of equation (3a), so in equation (4) t refers to abscissas and $\phi(t)$ refers to ordinates. The relation between the two systems of coördinates is given by

$$(5) \quad x = t\sigma + \bar{x}$$

for abscissas, and

$$(6) \quad y = \frac{N}{\sigma} \phi(t)$$

for ordinates. Equation (6) follows from (3a) and (4). If the area under the curve is taken as unity, then $y = \frac{1}{\sigma} \phi(t)$, that is, $\phi(t) = \sigma y$.

This says that since the abscissas are compressed by σ in changing from arbitrary units into standard units, so the ordinates must be stretched by σ if the area under the curve is to be the same in the two scales of measurement.

5. Tables of Standard Ordinates and Areas. One of the reasons for writing the equation in standard form is that the ordinates and

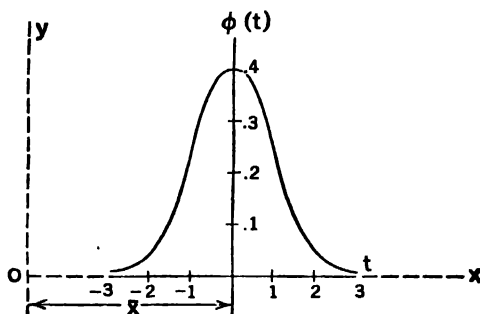


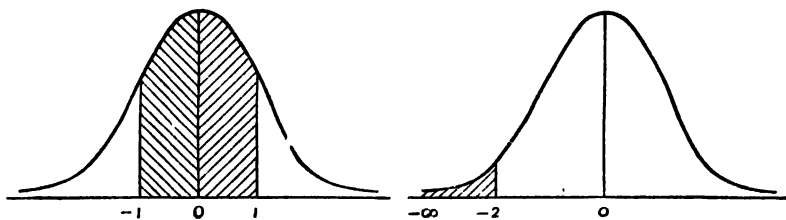
FIG. 22

areas may be tabulated once and for all. These tables are given in the Appendix. We see from (4) that $\phi(-t) = \phi(+t)$, i.e., the ordinates for negative values of t are the same as for the corresponding positive values of t , and the curve is symmetrical about the ordinate at $t = 0$. Therefore it is necessary to tabulate values of $\phi(t)$ for positive t 's only. Equation (4) may be graphed by plotting the points corresponding to a few well chosen values from the tables and drawing a smooth curve through them. (Figure 22.)

The curve approaches very close to the horizontal axis at each extremity but is asymptotic, that is, it does not quite touch the axis

no matter how far extended. We say its limits are at $-\infty$ and $+\infty$. Although the infinite abscissal range is never met in practice it may be characteristic of the "universe" from which a given distribution is a sample. Therefore, this infinite feature is useful in theoretical investigations. Moreover, even in representing observed distributions the infinite range causes no practical difficulty because the curve comes down to the horizontal axis very rapidly beyond $t = \pm 3$. The combined area at each extremity beyond $t = \pm 3$ is only .27 of 1% of the total area under the curve.

Partial areas between ordinates erected at various values of t , say between $t = a$ and $t = b$, are denoted by \int_a^b . Thus the area from



$t = 0$ to $t = 1$ is given by $\int_0^1 = .3413$. (See Table I, Appendix.)

Since the total area under $\phi(t)$ is taken as unity the area on either side of $t = 0$ is 0.5 and it is only necessary to tabulate the areas \int_0^t for positive values of t . Thus the area from $t = -1$ to $t = 0$ is equal to the area from $t = 0$ to $t = 1$. In symbols this would be stated as follows:

$$\int_{-1}^0 = \int_0^1.$$

Any other areas required may be found by an appropriate addition or subtraction of tabular values. For example, suppose the area below $t = -2$ is required. This is denoted by $\int_{-\infty}^{-2}$. Now the area from $-\infty$ to -2 equals 0.5 minus the area from -2 to 0. And the area from -2 to 0 is the same as from 0 to 2. That is,

$$\int_{-\infty}^{-2} = .5 - \int_{-2}^0. \quad \text{But} \quad \int_{-2}^0 = \int_0^2 = .4772.$$

$$\therefore \int_{-\infty}^{-2} = .5 - .4772 = .0228.$$

Both areas and ordinates for decimal values of t between tenths may be approximated by interpolating between the values given in the tables.

The illustrative examples following §6 will help the student become familiar with the tables. He should verify the answers and draw a simple sketch of the curve showing the ordinates or areas in each case.

The symbol $\int_{-\infty}^t$ denotes a cumulative relative frequency, *i.e.* the percentage of the total frequency N which is less than t . In order to find values of $\int_{-\infty}^t$ from the tables, for assigned values of t , the student should observe (from a figure) that

$$\int_{-\infty}^t = .5 \pm \int_0^t$$

the plus or minus sign to be used according as t is positive or negative.

6. Properties. A knowledge of the properties of the normal curve is essential for an intelligent use of the curve in practical statistics. A demonstration of some of these properties is beyond the scope of the present discussion although quite simple in the calculus. The following properties are the most important and interesting.

1. The mean, median, and mode coincide at $t = 0$. The height of the maximum ordinate in standard form is $1/\sqrt{2\pi}$ because when $t = 0$, $\phi(t) = 1/\sqrt{2\pi} = .3989$.

2. Since the standard deviation is the unit of measurement along the horizontal axis, $\sigma_x = 1$ in the t scale. Any t value may be converted into the corresponding x value by (5). In the vertical direction N/σ is the unit of measurement and any $\phi(t)$ ordinate may be converted into y units by means of (6).

The area under (3) in the range from $x = c$ to $x = d$ is denoted by

$$\int_c^d y \, dx.$$

If $t = a$ and $t = b$ denote the corresponding range in standard units, then

$$(7) \quad \int_a^b \phi(t) \, dt$$

denotes the corresponding area, in standard units, under (4). It is

shown in the calculus that $dx = \sigma_x dt$. Therefore from (6) we have

$$(8) \quad \int_c^d y dx = N \int_a^b \phi(t) dt.$$

If the interval goes from $x = c$ to $x = d$, (8) says that

$$(9) \quad \text{Frequency over } (c, d) = N \int_a^b$$

where

$$(10) \quad a = (c - \bar{x})/\sigma_x, \quad b = (d - \bar{x})/\sigma_x.$$

This merely means that the percentages (relative frequencies) obtained from the tables may be converted into numbers (frequencies) by multiplying the percentages by N .

3. The curve changes from concave to convex at $t = \pm 1$. In the x -scale, referred to the origin of x , these points are at $x = \bar{x} \pm \sigma_x$. They are called points of inflection and their position is important in making an accurate drawing of the curve.

4. The standard deviation is approximately 25% greater than the mean deviation. More precisely, $MD = \sigma \sqrt{\frac{2}{\pi}} = .798\sigma$. $\left(\sqrt{\frac{\pi}{2}} = 1.2533.\right)$

5. The quartiles, Q_1 and Q_3 , are equidistant from $t = 0$ and therefore from the mean. By definition

$$Q_3 \text{ is that value of } t \text{ for which } \int_{-\infty}^t = .75,$$

i.e., for which $\int_0^t = .25$. From the tables this is $t = .6745$. Therefore in arbitrary units,

$$Q_3 = \bar{x} + .6745\sigma_x \text{ and } Q_1 = \bar{x} - .6745\sigma_x.$$

6. The quartile deviation (semi-interquartile range) for a normal distribution will be denoted by E . Its value is

$$E = Q_3 - Q_1 = \frac{(\bar{x} + .6745\sigma) - (\bar{x} - .6745\sigma)}{2} = .6745\sigma.$$

In standard units this is $s = E/\sigma = .6745$.

7. The quantity E (or s) has a significance in probability theory. If a variable x is distributed according to the normal curve, the

probability is one half that a variate selected at random will have a value between $\bar{x} - E$ and $\bar{x} + E$. The reason for this statement is that 50% of the variates have values within this range. E is commonly, though somewhat ambiguously, called "probable error."

8. E is in units of x whereas s is a value of t , that is, s is the value $t = .6745$, and E is the value $x = .6745\sigma_x$. Just as σ_x may be used as a yardstick in scaling off a distribution on either side of the mean (§6, Chapter V), so may E or s be used in a similar manner. When thinking of them in this way it is useful to regard E as a yardstick about two-thirds the length of σ_x . The following table gives the end-points of certain intervals in t , x' , and x units, respectively, where $t = x'/\sigma_x$ and $x' = x - \bar{x}$.

End Points of Certain Intervals in t , x' , x

When σ is the unit			When E is the unit		
t	x'	x	t	x'	x
0	0	\bar{x}	0	0	\bar{x}
± 1	$\pm \sigma$	$\bar{x} \pm \sigma$	$\pm .6745$	$\pm .6745\sigma$	$\bar{x} \pm .6745\sigma$
± 2	$\pm 2\sigma$	$\bar{x} \pm 2\sigma$	± 1.349	$\pm 1.349\sigma$	$\bar{x} \pm 1.349\sigma$
± 3	$\pm 3\sigma$	$\bar{x} \pm 3\sigma$	± 2.023	$\pm 2.023\sigma$	$\bar{x} \pm 2.023\sigma$

The percentage distribution of area under the normal curve is given (approximately) in Figure 23 where σ_x is the unit of measurement along the horizontal axes and in Figure 24 where s is the unit. The percentages given in the figures may be regarded as abridged tables. Of course the tables in the Appendix will ordinarily be used in problems.

With reference to Figure 23, it is sometimes said that if values of x are normally distributed, the probability that a value chosen at random will fall within the range $x_1 \leq x \leq x_2$, where $x_1 = \bar{x} - \sigma_x$ and $x_2 = \bar{x} + \sigma_x$, is .68.

9. Astronomers and physicists have called h the "modulus of precision." From the relation $h = 1/(\sqrt{2}\sigma)$, it is evident that h increases as σ decreases. And as h increases, the curve (with N and m kept constant) becomes narrower in the neighborhood of m and in this sense h measures the closeness of the values of x to their mean.

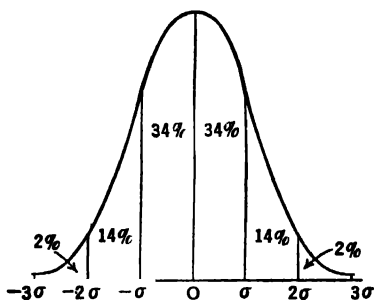


FIG. 23

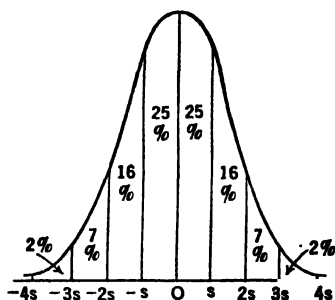


FIG. 24

10. The curve is symmetrical and $\alpha_3 = 0$. The fourth moment about the mean is equal to three times the square of the second moment about the mean, i.e., $\mu_4 = 3\mu_2^2$ and therefore $\alpha_4 = \mu_4/\mu_2^2 = 3$.

Examples

1. Find the ordinates of $\phi(t)$ for (a) $t = 2.3$, (b) $t = -2.3$, (c) $t = .67$.

Solutions from the tables in the Appendix:

- (a) $\phi(2.3) = .02833$
 (b) $\phi(-2.3) = .02833$
 (c) $\phi(.67) = .31874$

2. Find the following areas under $\phi(t)$ and use the integral notation:

- (a) From $t = 0$ to $t = 3.00$
 (b) From $t = 1.5$ to $t = 2.5$
 (c) From $t = -2$ to $t = 1.3$
 (d) From $t = 0$ to $t = 0.6745$

Solutions from the tables:

- (a) The required area is given by \int_0^3 which we find to be .49865.

- (b) The area from $t = 0$ to $t = 1.5$ is $\int_0^{1.5} = .43319$, and from $t = 0$ to $t = 2.5$ is $\int_0^{2.5} = .49379$. Therefore the required area is $\int_{1.5}^{2.5} = \int_0^{2.5} - \int_0^{1.5} = .0606$.

- (c) Since the area from $t = 0$ to $t = -2$ is the same as from $t = 0$ to $t = +2$ we have

$$\int_{-2}^{1.3} = \int_0^2 + \int_0^{1.3} = .47725 + .40320 = .88045.$$

(d) Here we must interpolate:

$$\text{For } t = .67, \quad \int_0^t = .24857$$

$$\text{For } t = .6745, \quad \int_0^t = A \text{ (say)}$$

$$\text{For } t = .68, \quad \int_0^t = .25175.$$

Therefore

$$\frac{A - .24857}{.25175 - .24857} = \frac{.0045}{.01}$$

whence

$$A = .25.$$

3. Show that for equation (3), the percentages of area outside the given ranges are as stated below:

$$\text{Above } \bar{x} + \sigma = 15.87\%$$

$$\text{Outside } \bar{x} \pm \sigma = 31.74\%$$

$$\text{Outside } \bar{x} \pm 2\sigma = 4.56\%$$

$$\text{Outside } \bar{x} \pm 3\sigma = 0.27\%$$

Solution: Converting these ranges into t units, and remembering that only the positive half of the area under $\phi(t)$ is tabulated and equals .5, we have

$$\begin{aligned} \text{Area above } t = 1 \text{ is } .5 - \int_0^1 &= .1587 \\ &= 15.87\% \end{aligned}$$

$$\text{Area outside } t = \pm 1 \text{ is } 2(15.87\%) = 31.74\%$$

$$\begin{aligned} \text{Area outside } t = \pm 2 \text{ is } 2\left(.5 - \int_0^2\right) &= .0456 \\ &= 4.56\% \end{aligned}$$

$$\begin{aligned} \text{Area outside } t = \pm 3 \text{ is } 2\left(.5 - \int_0^3\right) &= .0027 \\ &= 0.27\% \end{aligned}$$

4. Given $N = 1500$, $\bar{x} = 75$, $\sigma_x = 10$. If the variates are distributed according to the normal curve, (a) find the value of x for which $\text{cum } f = 800$, (b) for which $\text{cum } f = 450$, (c) how many of the N variates lie where $x < 80$?

Solutions:

$$(a) \text{ By definition, } \text{cum } f = \int_{-\infty}^x$$

$$\text{and from (8), } \int_{-\infty}^x = N \int_{-\infty}^t$$

$$\therefore 800 = 1500 \int_{-\infty}^t$$

$$\text{i.e., } \int_{-\infty}^t = .5333.$$

But
$$\int_{-\infty}^t = .5 + \int_0^t$$
$$\therefore \int_0^t = .0333$$

whence from the tables,

$$t = .083.$$

Substituting in equation (5),

$$x = 75.83.$$

(b) We have
$$\int_{-\infty}^t = 45/150 = .3 \text{ and } t \text{ is negative.}$$

Since
$$\int_{-\infty}^t = .5 - \int_0^t$$

we have
$$\int_0^t = .2$$

whence we find that

$$t = -.524$$

so

$$x = 69.76.$$

(c) From the relation $t = (x - \bar{x})/\sigma_x$ we find that

$$t = .5 \text{ when } x = 80.$$

From the tables,

$$\int_{-\infty}^{.5} = .69146.$$

From (8) we have

$$\begin{aligned} \int_{-\infty}^x &= 1500 (.69146) \\ &= 1037.2. \end{aligned}$$

Exercises

1. Find $\phi(2.65)$, $\phi(-1.46)$, $\phi(0)$.
2. Find t if $\phi(t) = .1257$, $.0325$, $.0034$, respectively.
3. Find the following areas under $\phi(t)$, and draw a figure in each case:

(a) \int_0^{∞} , $\int_{-1.2}^{1.2}$, $\int_{-\infty}^{1.2}$, $\int_{1.2}^{\infty}$, $\int_{-1.2}^{\infty}$.

(b) $\int_{-.37}^{.37}$, $\int_{-.6745}^{.6745}$.

4. Find t , given the partial areas:

$$2 \int_0^t = .5, \int_0^t = .27457, \int_{-t}^t = .999730.$$

5. Verify the percentages given in Figures 23 and 24.
6. (a) How far from the median of a normal distribution is the first quartile?
(b) In a certain normal distribution $\bar{x} = 89$ and $Q_1 = 75.51$. What is σ_x ?
7. For a normal distribution: $N = 1000$, $\bar{x} = 20$, $\sigma_x = 2$.
(a) What is E ?
(b) Find the value of Q_3 .
(c) What values of x will include the middle 500?
(d) The middle 75%?
8. If $N = 300$, $\bar{x} = 75$, $\sigma_x = 15$, for a normal distribution:
(a) What is the value of the first quartile?
(b) The third quartile?
(c) How many variates are between $x = 60$ and $x = 90$?
9. In a college the 8 grades A, A-; B, B-; C, C-; D, and F are given. On the assumption that mathematical ability is normally distributed, how many out of a total of 1000 should receive each grade? Assume that \bar{x} is the boundary between the C and B- grades and that each grade interval is $.8\sigma$. What range in standard units on either side of \bar{x} is thereby assumed to include all the grades?
10. What are the percentages of a normal distribution outside $\bar{x} \pm t\sigma$ for $t = 1, 2, 3$?

7. Curve Fitting. It should be remembered that a set of data collected and presented in the form of a frequency distribution is merely a sample of a general type called its universe. Other samples from that universe might yield somewhat different frequency distributions.

For certain purposes it may be desirable to fit a normal curve to a unimodal distribution which is reasonably symmetrical and appears to be of the normal type. The theoretical curve idealizes the recalcitrant observational data and smooths out the irregularities due to sampling fluctuations.

In fitting equation (3a) to a given distribution, we assume that

(1) *The given frequency N represented by a histogram equals the area under the curve, and*

(2) *The mean and standard deviation of the observed distribution equal, respectively, the mean and standard deviation of the theoretical distribution represented by the curve.*

A normal curve is a mathematical model of a hypothetical universe. In identifying such a universe with (3a) only its form is specified by the model. The parameters are (usually) unknown. An estimate of a parameter by the use of an appropriate function of the observed data is called a *statistic*. Assumption (2) above means,

then, that we replace each of the parameters by the corresponding statistic.¹

The procedure of fitting a normal curve to an observed distribution will now be illustrated with the data of Table 21, p. 88. We substitute

$$\begin{aligned}\bar{x} &= 47.712 \\ \sigma_x &= 5.772 \\ N &= 1000\end{aligned}$$

in equation (3), and obtain

$$y = \frac{1000}{5.772 \sqrt{2\pi}} e^{-\frac{(x-47.712)^2}{2(5.772)^2}}.$$

To make use of a table of standard ordinates in graphing this equation we transform it into standard units by setting

$$(a) \quad t = \frac{x - 47.712}{5.772} = .17325x - 8.2661$$

and write

$$(b) \quad y = \frac{N}{\sigma} \phi(t) = 173.25\phi(t).$$

Appropriate values to assign x in equation (a) are the *end- x* and *mid- x* values of the given distribution. The use of a computing machine in changing x values into corresponding t values is explained in §6, Chapter IV. Thus we obtain the values in the second column of Table 29. We may then enter the table in the Appendix for the corresponding ordinates, $\phi(t)$. These are converted into y values by equation (b). The curve may then be drawn by plotting

¹ It is shown in Part II that a better estimate of the variance in the universe is obtained by multiplying the variance of the observed distribution by $N/(N-1)$. Because of this fact some writers, denoting this result by s^2 , define the variance of an observed distribution by

$$s^2 = \frac{1}{N-1} \sum_1^k f_i(x_i - \bar{x})^2.$$

The distinction between the two definitions is not an important one, in the author's opinion, for beginning students who are learning the descriptive methodology of statistics. And in curve fitting, the numerical difference is negligible because N is fairly large. The distinction is important, however, in the theory of small samples (Part II).

TABLE 29. $t = .17325x - 8.2661$, $y = 173.25\phi(t)$

x	t	$\phi(t)$	y	f/c
27.5	-3.502	.00086	0.15	
29.5	-3.155	.00275	0.48	0.25
31.5	-2.809	.00772	1.34	
33.5	-2.462	.01927	3.34	3.50
35.5	-2.116	.04253	7.37	
37.5	-1.769	.08344	14.46	14.00
39.5	-1.423	.14494	25.11	
41.5	-1.076	.22361	38.74	43.00
43.5	-0.730	.30563	52.95	
45.5	-0.383	.37072	64.23	61.25
47.5	-0.037	.39866	69.07	
49.5	0.310	.38023	65.87	65.75
51.5	0.656	.32230	55.84	
53.5	1.003	.24124	41.79	39.00
55.5	1.349	.16060	27.82	
57.5	1.696	.09469	16.41	16.75
59.5	2.042	.04960	8.59	
61.5	2.389	.02299	3.98	5.75
63.5	2.735	.00948	1.64	
65.5	3.082	.00346	0.60	0.75
67.5	3.428	.00111	0.19	

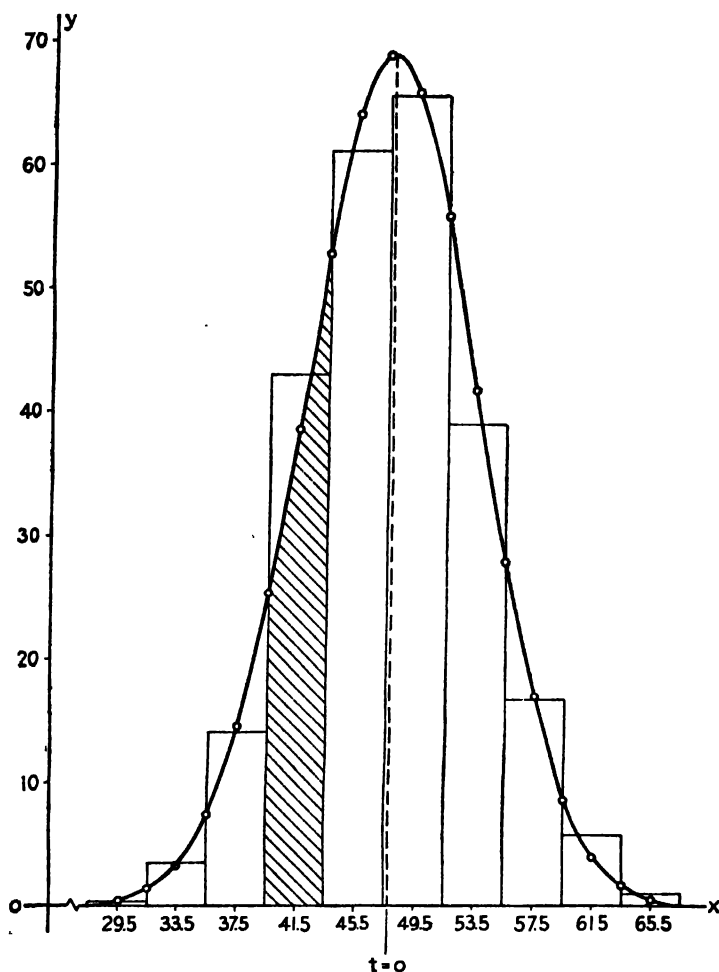
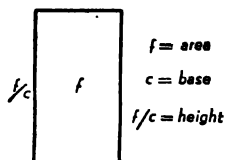


FIG. 25 — NORMAL CURVE FITTED TO HISTOGRAM REPRESENTING WEIGHT DISTRIBUTION OF GLASGOW SCHOOLGIRLS (TABLE 21)

The smooth curve is plotted from the points (x, y) given in Table 29. The column headed f/c in that table gives the heights of the rectangles in the histogram, $c = 4$. When both the curve and the histogram are to be drawn, it is best to draw the curve first so that the presence of the histogram will not prejudice one into trying to make the curve fit the histogram.

the x and y values. (Figure 25.) The curve should be drawn so as to be symmetrical with respect to the ordinate at the mean and its points of inflection should be at a distance from the mean equal



to σ . The student should observe that every pair of (x, y) values computed in Table 29 furnishes two points for the graph, each symmetrical to the other with respect to the mean ordinate. Both points should be used in drawing the curve but only the computed points should be left permanently in the graph.

After the curve is drawn, the histogram for the observed data may be constructed. The column headed f/c gives the heights of the rectangles on the same scale as the ordinates of the curve.

8. Graduation. The areas under the fitted curve and over the class intervals are called theoretical frequencies. Thus in Figure 25 the shaded area represents the theoretical frequency corresponding to the observed frequency which is represented by the rectangle the mid-point of whose base is 41.5 pounds. The determination of the theoretical frequencies is called "graduation by the normal curve." It is a process of smoothing out the data to fit the curve. The method is shown in Table 30 for the data represented by Figure 25.

In order to enter a table of standard areas we must change the *end- x* values into t values. These are given in the third column of Table 30. They are part of the values already computed for Table 29.

The entries in the column headed $A = \int_{-\infty}^t$ are the $(\text{cum } f)/N$ values of the standard curve for the given end-points. The entries in the column headed ΔA are obtained by differencing the preceding column. (See last paragraph of §9, Chapter I.) They are the percentages $p = f/N = \Delta A$ to be expected in the various intervals on the hypothesis of a normal distribution. Therefore $N\Delta A$ gives the *numbers* to be expected, that is, the theoretical frequencies.

The student should study this table until he becomes familiar with all the operations involved and what they mean. He should distinguish between the purposes of Tables 29 and 30.

9. Purpose of a Graduation. If, for the distribution of graduated frequencies, the mean, standard deviation, and total frequency are found, their values will be precisely those of the corresponding moments in the observed frequency distribution. This must be so, because these were the conditions imposed in the process of gradu-

TABLE 30

<i>Observed Frequency</i>	<i>Boundary x</i>	<i>t</i>	<i>$A = \int_{-\infty}^t$</i>	ΔA	<i>$N\Delta A =$ Theoretical Frequency</i>
	$-\infty$	$-\infty$.0000		
1				.0025	2.5
	31.5	-2.809	.0025		
14				.0147	14.7
	35.5	-2.116	.0172		
56				.0602	60.2
	39.5	-1.423	.0774		
172				.1553	155.3
	43.5	-0.730	.2327		
245				.2527	252.7
	47.5	-0.037	.4854		
263				.2587	258.7
	51.5	0.656	.7441		
156				.1674	167.4
	55.5	1.350	.9115		
67				.0679	67.9
	59.5	2.042	.9794		
23				.0175	17.5
	63.5	2.735	.9969		
3				.0031	3.1
	∞	∞	1.000		
Totals				1.0000	1000.0

ation. Moreover, the observed values of skewness and kurtosis as given by α_3 and α_4 will not differ appreciably from the theoretical values if the fitting of the normal curve to the observed distribution was justified.

Since the above parameters characterize a distribution, the observing student may wonder why a distribution should be graduated if the values of these constants are unaltered in the process.

There are three main reasons why a student should be taught to graduate a curve. The first, and least important, has to do with the use of a smooth curve in place of a jagged sample. The second, and most important, is that it is necessary for the mathematical development of statistics that the mathematician should be told what assumptions he may make. These usually depend on the types of frequency curves which can be depended on to fit phenomena. . . . A third reason, intermediate in importance between the other two, is that in testing *a priori* theories in various fields, it is often necessary to test the efficacy of the frequency distributions which are results of these theories.¹

The second and third of the above reasons may seem somewhat abstruse, but it is not easy to give completely satisfactory explanations of them at this level of exposition. About all we can say at this time is that the distribution of variation of a variable x about its mean value is a fundamental statistical concept and in certain theoretical investigations it is very important that we have mathematical functions which are capable of representing such distributions. This is particularly true in sampling theory which will be discussed in Part II.

The first reason is more readily understood. Occasionally in practical problems it may be desirable to use the theoretical frequencies obtained by graduation in place of the observed data which probably contain irregularities due in part to grouping, in part to sampling fluctuations. We cite here two illustrations.

Example 1. A company which operates a chain of men's haberdashery stores planned to bring out a new line of about 100,000 light weight sport shirts suitable for camping, hunting, etc. The question arose as to the determination of the number of each size that should be ordered from the factory. Their previous distribution of sizes had not been satisfactory because the demand for certain sizes had been different from the number manufactured. Therefore the statistical department was requested to recommend the distribution of the proposed order according to neck sizes. The solution of the problem hinged upon the availability of data giving the measurements of neck circumferences of a large sample of men. Satisfactory data were found in the "Reports of the Medical Department of the United States Army in the World War," which gave a table of the

¹ *Journal of the American Statistical Assoc.*, vol. XXVI, March 1931, Supplement, p. 36.

neck measurements in centimeters of 95,102 white troops at demobilization. Since these data are tabulated in class intervals which are slightly different from the ranges used in standard shirt-band sizes, a slight adjustment was necessary. But essentially a normal curve was fitted to this distribution and the graduated frequencies were taken as the number of potential customers for each shirt size. The result was quite satisfactory.

Example 2. A well known and interesting illustration of the desirability of smoothing occurs in the census returns. The census takers' records show more persons alive at age 30 than at age 29, more at age 35 than at age 34, more at 40 than at 39, etc. This is probably due to the fact that men (as well as women) do not tell their exact ages. A person who is actually 41 or 42 and known to be 40 or so, says he is 40. The recorded data show artificial bumps at every age which is a multiple of 5. Naturally the Census Bureau prefers the smoothed results to be observed. The student should not infer that the curve used to smooth these data is the normal type. The "life curve" is a continuously decreasing function. However, the same kind of quinquennial irregularity occurs in other actuarial data which do approximate the form of a normal curve. Many examples are given in Elderton, *Frequency Curves and Correlation*.

10. Probability. A frequency curve is sometimes called a probability curve. The link connecting frequencies with probabilities has its starting point in the following definition:

DEFINITION. *If out of N mutually exclusive and equally likely events, f are distinguished by some property A , the probability of an event bearing the property A is f/N .*

The definition implies that probability is measured by a number in the range 0 to 1, the lower limit denoting impossibility and the upper limit denoting certainty.

Since the total area under the curve represented by (4) is unity, any partial area denoted by (7) can be interpreted as the probability that a value of t selected at random from a normal distribution (4) lies between $t = a$ and $t = b$.

Example 1. Refer to the data of Table 8, Chapter I. Let us assume that a normal curve was fitted to this distribution and that the fit seemed (by visual inspection) to be reasonably good. Generalizing on the experience shown in the table, the telephone company wishes to estimate the probability that a call (of the same type of message as that in the table) will be between (say) 500 seconds and 600 seconds in length.

Solution. Using (10),

$$a = (500 - 477.3)/148.5 = 0.15,$$

$$b = (600 - 477.3)/148.5 = 0.83.$$

Under the implicit assumptions, the required probability is $P = \int_{0.15}^{0.83} = 0.24$.

Example 2. Referring to Example 1 above, find the probability that the length of a telephone call will differ numerically from the mean of the table by as much as 5 minutes.

Solution. We find $|t| = 300/148.5 = 2.02$. The probability of a deviation not greater, numerically, than 300 seconds is $P = 2 \int_0^{2.02} = 0.96$, approximately. Then the probability of a numerical deviation as large as (or larger than) 300 seconds is $Q = 1 - P = 0.04$. This would be represented graphically by the area under the curve *outside* $t = \pm 2.02$.

11. Probability Paper. The cumulative frequencies for the normal $\phi(t)$ curve are given by $A = \int_{-\infty}^t$. As t varies from $-\infty$ to $+\infty$, A varies from 0 to 1, and for the finite range $t = \pm 3$ (commonly met in practice) A varies from 0.00135 to 0.99865. (Verify.) Regarding A as a function of t , values of (t, A) from the tables may be plotted and the resulting points joined by a smooth curve.

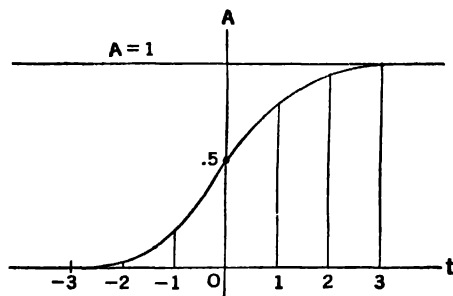


FIG. 26 -- OGIVE OF THE NORMAL CURVE

When graphed on an algebraic scale this curve is the ogive of the normal curve. It is also called the integral curve of $\phi(t)$. As indicated in Figure 26, the ordinate of the ogive is zero at $t = -\infty$, .5 at $t = 0$, and the ogive approaches the line $A = 1$ asymptotically.

Now imagine the vertical scale of Figure 26 stretched in such a way that the ogive becomes a straight line. The stretching required will be greatest around the line $A = 0.5$ and gradually diminish as the distance from this line increases.

Paper so ruled that the (t, A) graph is a straight line is called *probability paper*. It is readily obtainable¹ and is convenient for many purposes. Thus, by plotting *cum f* for an observed distribution on probability paper, one may observe how closely it approxi-

¹ The Codex Book Company, New York.

mates a straight line and hence get an idea of how nearly normal it is. One may thus locate graphically the median, quartiles, etc., and estimate frequencies between given limits.

A more complete discussion giving references to writers who suggested and developed the use of probability paper may be found in the *Journal of the American Statistical Association*, vol. XXVI, June 1931, p. 178.

Exercises

1. Construct three normal curves on the same axes according to the following specifications. Compute ordinates at intervals of $.5\sigma$ from the mean in the range $\bar{x} \pm 3\sigma$.

Curve	σ_x	\bar{x}	N
A	10	50	400
B	10	50	800
C	10	50	1200

Suggested form for computations:

x	t	$\phi(t)$	y		
			A	B	C
20	-3				
—	—				
—	—				
—	—				
80	3				

2. Construct three normal curves on the same axes according to the following specifications. Compute ordinates at intervals of $.5\sigma$ from the mean.

Curve	σ_x	\bar{x}	N
A	15	50	1000
B	10	50	1000
C	5	50	1000

Suggestion:

x			t	$\phi(t)$	y		
A	B	C			A	B	C
5	20	35	-3				
—	—	—	—				
—	—	—	—				
—	—	—	—				

Observe that:

$$y_C = 200\phi(t)$$

$$y_B = \frac{1}{2}y_C = 100\phi(t)$$

$$y_A = \frac{1}{3}y_C = \frac{200}{3}\phi(t).$$

3. Verify the entries in Tables 29 and 30.
4. For the following distribution:
 - (a) Find the equation of the best fitting normal curve, and plot the curve and histogram.
 - (b) Find the graduated frequencies.

<i>mid-x</i>	2	4	6	8	10
<i>f</i>	1	4	6	4	1

5. Graduate the distribution in Table 8, §11, Chapter I. Also find the ordinates of the best fitting normal curve and plot the curve and histogram.
6. A distribution of the weekly wages of 906 anthracite miners showed the following results:

$$\bar{x} = \$36.13$$

$$\alpha_3 = 0.007$$

$$\sigma_x = \$8.87$$

$$\alpha_4 = 3.02$$

Assuming a normal distribution, estimate the number of the 906 miners who received weekly wages (a) in excess of \$45, (b) less than \$25.

7. An urban electric railway company operating a large city subway uses thousands of electric light bulbs in its underground stations. On January 1, 1947, the company put into service 5000 new light bulbs. Let it be assumed that these 5000 bulbs will have a mean life of 50 days, a standard deviation of 19 days, and that their lives conform to the normal curve.

If January 1 is counted as a full day in the life of the bulbs: (a) How many bulbs out of the 5000 new ones would have had to be replaced by midnight January 31, 1947? (b) How many by March 10, 1947?
8. Which properties of the normal curve may be used as criteria in passing judgment on the normality of an observed distribution? Would you say that the distributions referred to in Table 23 are approximately normal?
9. Graph the ogive of the normal curve by plotting values of (t , A) in the range $t = \pm 3$, (a) on an algebraic scale, (b) on probability paper.
10. What famous mathematicians' names are associated with the normal curve? When did these men live? Which of them should most appropriately be credited with the discovery of this curve?
11. (Camp) The standard deviation of a certain set of 100,000 high school grades was 11%, and the mean grade was 78%. Assume the distribution to have been normal, and, being careful not to confuse percentage in the sense of grade with a percentage of frequency, answer the following questions: How many grades were (a) above 90%, (b) below 70%? (c) What

was the highest grade of the lowest 1000? (d) Within what limits did the middle 90,000 lie? (e) What was the semi-interquartile range?

12. (Camp) Answer all the questions of Exercise 11 with reference to a set of 100,000 grades in which the median was 83% and Q_3 was 90%. Also find σ_x .
13. In a certain normal distribution, $N = 1000$, $\bar{x} = 50$, $\sigma_x = 10$. For this distribution:
- (a) Convert the following x 's into the corresponding t 's,

x	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85
t															

- (b) Find from the tables the values of $\phi(t)$ for the t values in (a).
- (c) Convert the $\phi(t)$ values obtained in (a) into y values.
- (d) Plot the (x, y) values in (a) and (c) and draw a smooth curve through them.
- (e) Find the cumulative relative frequencies, $A = \int_{-\infty}^t$, for the values of t in (a).
- (f) Difference your results in (e) by finding ΔA .
- (g) Convert the percentages in (f) into frequencies.
- (h) Explain the meaning of your results in (g) with reference to the figure for (d).
- (i) Find the number of variates between $x = 42$ and $x = 74$.
- (j) Find the values of x for which $\text{cum } f = 250, 600, 750$, respectively.
14. Given a normal distribution in which $N = 800$, $\bar{x} = 40$, $\sigma_x = 7$. Find the numerical value of each of the following.

$$Q_1, Q_2, Q_3, E, N \int_{t=0}^{t=s}$$

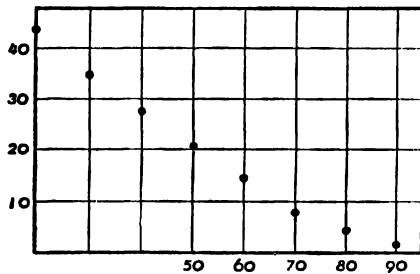
15. Suppose $N = 5000$ variates are normally distributed such that $\bar{x} = 50$ and $E = 13.49$. Without using the tables find the value of the following: quartiles, median, mode, standard deviation, mean deviation, x for which $\text{cum } f = 1250$.
16. Suppose there are N values of a variable v which are normally distributed with mean = 0 and variance = 25.
- (a) Give the equation of the curve which represents the distribution.
- (b) If there are 793 values between $v = -5$ and $v = 0$, determine N .
- (c) What percent of N have values larger than $v = 10$?
- (d) Determine the value of v for which $\text{cum } f = .75N$.

CHAPTER VII

CURVE FITTING

1. Empirical Expressions. The preceding chapters have dealt with the description and characterization of frequency distributions. We have considered three general methods of description: (1) graphical devices, (2) the method involving calculation of averages and measures of dispersion, (3) the method which is sometimes called *analytical*. This latter method consists in describing the distribution by an equation, and we considered only one such analytical expression, the normal curve.

Example 1. Expectation of Life¹ at various ages.



Age	Expectation
20	42.20
30	35.33
40	28.18
50	20.91
60	14.10
70	8.48
80	4.39
90	1.42

However, another branch of statistics is concerned with data which may not be classed under frequency distributions, but which may be described by simple equations.

When one variable is a function of another in applied mathematics the mathematical relation between them is not always known. As we mentioned in Chapter II, the only information regarding this functional relationship may be a set of pairs of values obtained by experimental or observational means. These pairs of values may be regarded as coördinates of points and plotted. In doing so, the values

of the variable which is regarded as independent are taken as abscissas, and those of the dependent variable as ordinates.

The general problem in such cases is to find, if possible, an analytic

¹ By expectation of life at any age is meant the average number of years lived by persons attaining that age, as given in the *American Experience Mortality Table*.

expression of the form $y = f(x)$ for the functional relationship suggested by the data. Equations obtained to fit observed data as well as possible are called empirical to distinguish them from the rational expressions of pure mathematics which can be derived from reasoning. This general problem is called *curve fitting*. It is also sometimes referred to as "smoothing" the given data.

We will consider three types of functions: *linear*, *quadratic*, and *exponential*.

2. Linear Functions. We know from algebra that the general form of a linear equation in two variables is

$$Ax + By = C$$

where A , B , and C are arbitrary constants.

When $B \neq 0$, the equation may be solved for y , giving $y = -(A/B)x + C/B$ which is of the form

$$(1) \quad y = mx + k$$

and which is the form we will ordinarily use to represent a straight line.

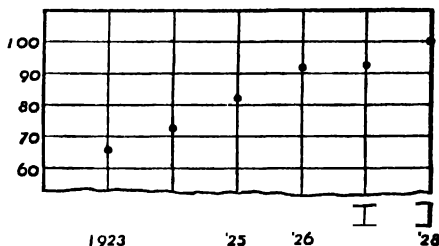
The special cases where A or B or C are zero is as follows:

When $A = 0$, then $y = C/B$, which is of the form $y = k$. This is a line parallel to the x -axis. When $B = 0$, the equation takes the form $x = k$ which is a line parallel to the y -axis. When $C = 0$, then $Ax + By = 0$ which is a line passing through the origin.

The graph of (1) is a straight line (which explains the term "linear"). A characteristic property of a linear function is revealed at once by its graph. This is the fact that the ratio of a change in y to the corresponding change in x is *constant*. Thus, if two points (x_1, y_1) and (x_2, y_2) are chosen on the line, the value of the ratio

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

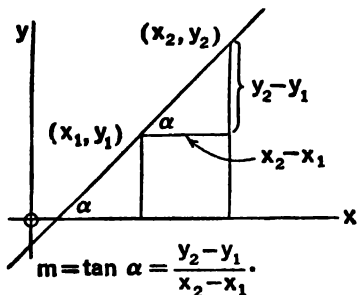
Example 2. Yearly Production of Cigarettes in the United States



Year	Billions
1923	66.7
1924	72.7
1925	82.3
1926	92.1
1927	93.0
1928	100.0

is independent of the points chosen. This ratio gives the average rate of change of any function over the interval $\Delta x = x_2 - x_1$. In the case of a linear function, m defines the *rate of change* of the function.

Graphically, m is the *slope* of the line. It is the tangent of the *angle of inclination* α (alpha) which the line makes with the positive x -axis.¹ Lines having the same slope are parallel, and conversely.



It is shown in analytic geometry that we may obtain the slope of a straight line from its equation if we solve for y and take the coefficient of x . Thus in $2x - y = 5$, $y = 2x - 5$ and the slope is 2.

Conversely, if we know the slope of a line and the coordinates of any point on the line we can write its equation from the relation

$$(2) \quad y - y_1 = m(x - x_1)$$

which is called the *point-slope* form of a straight line. Thus, given that $(2, -1)$ is a point on a line whose slope is 2, the equation of the line is therefore $y + 1 = 2(x - 2)$ or $2x - y = 5$.

Or again, remembering that m is defined by a ratio involving the coordinates of two points on a line, we can obtain the equation of a line if we know any two points which lie on it. From the definition of m and (2), we have

$$(3) \quad y - y_1 = \frac{y_2 - y_1}{x_2 - x_1} (x - x_1)$$

which is known as the *two-point* form of a straight line. Thus, given that $(2, -1)$ and $(6, 7)$ are two points on a line, its equation is

$$y + 1 = \frac{7 + 1}{6 - 2} (x - 2) \quad \text{or} \quad 2x - y = 5.$$

3. Quadratic Function. A quadratic function of a variable v is a polynomial of the second degree in v which may be expressed in the form $Av^2 + 2Bv + C$ where A , B , and C are fixed real numbers.

¹ When the line is vertical, $\alpha = 90^\circ$ and m does not exist. Then $\Delta x = 0$ and division by zero is excluded in our algebra.

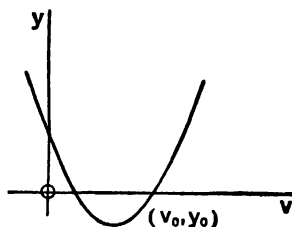
The minimum value of such a function is useful in statistics. We have

$$\begin{aligned} Av^2 + 2Bv + C &= \frac{1}{A} [Av^2 + 2ABv + AC] \\ &= \frac{1}{A} [(Av + B)^2 + (AC - B^2)]. \end{aligned}$$

Since $(Av + B)^2$ is positive or zero and $(AC - B^2)$ does not involve the variable, we have the following:

Theorem I. *If A is positive the minimum value of $Av^2 + 2Bv + C$ occurs when $Av + B = 0$; the minimum value is $(AC - B^2)/A$.*

The graph of the equation $y = Av^2 + 2Bv + C$, ($A > 0$), is a parabola which opens upward and whose vertex is where $v = -B/A$. Of course the function has its minimum value at this vertex, viz.: (v_0, y_0) where $v_0 = -B/A$, $y_0 = (AC - B^2)/A$.



Exercises

- (Wilson and Tracy) The premium (\$ y) on a \$1000 life insurance policy for various ages (x years) is given in the following table. Draw a graph exhibiting y as a function of x . Estimate from the graph the premium at age 32 and at age 43; also the age at which the premium is \$52.

x	20	25	30	35	40	45	50	55	60
y	18.78	21.02	23.86	27.54	32.36	38.83	47.68	59.88	76.94

- Find an equation of each of the lines through two points given as follows: (a) (2, 6), (4, 5); (b) (0, 3), (1, 6).
- Find the equation of a line through the point (2, 3) and parallel to the line $4x + 5y = 7$.
- (a) Find the value of x for which $f(x) = 2x^2 - 8x + 9$ has a minimum value. (b) What is this minimum value? (c) Draw a graph of $y = f(x)$ and show the meaning of your answers to (a) and (b).
- How would the theorem in §3 be affected if $A < 0$?
- Prove that the second moment of x is a minimum when taken about the mean of x .

Hints. Solution 1.

$$\begin{aligned}\text{Let } f(v) &= \frac{1}{N} \sum_1^N (x_i - v)^2 \\ &= v^2 - 2\bar{x}v + \frac{1}{N} \sum_1^N x_i^2.\end{aligned}$$

By the theorem of §3, show that $f(v)$ is a minimum when $v = \bar{x}$.

Solution 2. By definition,

$$\begin{aligned}\mu_2 &= \frac{1}{N} \sum_1^N (x_i - \bar{x})^2 \\ \nu_2 &= \frac{1}{N} \sum_1^N (x_i - v)^2, v \neq \bar{x}.\end{aligned}$$

Is $\mu_2 < \nu_2$?

Solution 3. for calculus students. From $f(v)$ as derived above,

$$f'(v) = -\frac{2}{N} \sum_1^N (x_i - v).$$

Set $f'(v) = 0$ and solve for v . Since $f''(v) > 0$, $v = \bar{x}$ yields a minimum, not a maximum.

7. Show that the value of k for which $f(k) = Nk^2 + 2k(m \sum_1^N x_i - \sum_1^N y_i) + C$ is a minimum is defined by

$$m \sum_1^N x_i + Nk = \sum_1^N y_i.$$

4. Fitting a Straight Line. The preceding discussion is intended as a basis for the presentation of certain methods of fitting a line to data. The equation $y = mx + k$ represents a family or set of lines corresponding to different values of the arbitrary constants m and k . As noted previously, such constants are called parameters. The process of finding the best fitting line for any given data consists in determining m and k . By "best fitting" we mean best under a criterion of approximation specified by a method. We will consider three such methods: (a) *graphical*, (b) *the method of moments of ordinates*, (c) *the method of least squares*.

5. Graphically. A straight line is drawn (preferably with the aid of a transparent ruler) to fit as closely as possible the plotted points. To find the equation of this line, select two points on the line and estimate their coördinates (x_1, y_1) and (x_2, y_2) . Substituting these coördinates in the "two-point" form of the line (3), we get the desired equation.

If the first point is chosen so that $x_1 = 0$ the numerical work of simplifying the equation is somewhat lessened.

Example 3. Fit a line graphically to the data in Example 2.

We take the origin of x at 1923, hence from the figure ($x_1 = 0, y_1 = 67$) and ($x_2 = 5, y_2 = 100$).

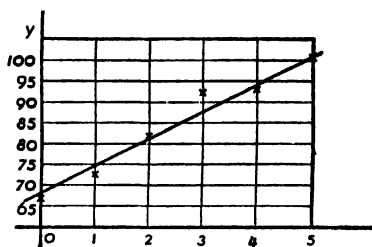
By equation (3),

$$y - 67 = \frac{100 - 67}{5} x$$

Therefore,

$$y = 6.6x + 67$$

is the required equation.



	x	y
(1923)	0	66.7
	1	72.7
	2	82.3
	3	92.1
	4	93.0
(1928)	5	100.6

The graphical method is open to the objection that it depends upon the judgment of the investigator. Different people will locate the line in different positions and therefore obtain different equations. However, where only approximate results are needed it is usually quite satisfactory.

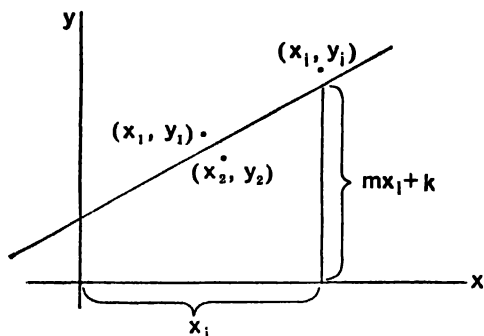
6. Method of Moments. In equation (1) y is not only a function of x but it is also a function of the *parameters* m and k . This functional relationship may be expressed symbolically by the notation $f(x, m, k)$. Given the functional form of a curve $y = f(x, a, b, c, \dots)$ the parameters a, b, c, \dots may be determined by obtaining expressions for as many moments of the *computed* or *functional* y 's as there are parameters in the function and equating these to the *numerical* moments of corresponding order of the observed or empirical y 's. A solution of the resulting equations, theoretically possible, gives the "best" values of the parameter. This is the *method of moments of ordinates*. For a set of N values of (x_i, y_i) the r th moment of y is defined by the expression

$$\frac{1}{N} \sum_{i=1}^N x_i^r y_i$$

where r is zero or a positive integer.

In fitting a straight line by this method we obtain two equations involving m and k if we equate the *zeroth* and first moments of the observed y 's to the *zeroth* and first moments, respectively, of the y 's

computed from the assumed equation $y = mx + k$. All moments are taken about the origin of x . These two equations may then be solved for m and k . The procedure will be made clear by the figure and explanation below.



x	${}_o y$	x	${}_c y$
x_1	y_1	x_1	$mx_1 + k$
x_2	y_2	x_2	$mx_2 + k$
\dots	\dots	\dots	\dots
x_i	y_i	x_i	$mx_i + k$
\dots	\dots	\dots	\dots
x_n	y_n	x_n	$mx_n + k$

Suppose we are given N pairs of values of x and y . Denote the given or observed y 's by ${}_o y$ and the computed y 's by ${}_c y$. For the observed y 's, the first moment is $\frac{1}{N} \sum x_i y_i$, and the zeroth moment is

$\frac{1}{N} \sum x_i {}_o y_i$. By a "computed y " corresponding to any value of x we mean the result obtained by substituting that value of x in the equation $y = mx + k$, and solving for y . Thus, for any value of x , say x_i , we obtain $mx_i + k$ for the corresponding computed y_i . Graphically, it is an ordinate of the line. Therefore, the first moment of the computed y 's is $\frac{1}{N} \sum x_i (mx_i + k)$, and the zeroth moment is

$\frac{1}{N} \sum x_i^0 (mx_i + k)$. Applying the principle of moments we have

$$\begin{array}{lcl} \text{zeroth moment} & \text{observed} & \text{computed} \\ & \sum y_i = & \sum (mx_i + k) \\ \text{first moment} & \sum x_i y_i = & \sum x_i (mx_i + k) \end{array}$$

where the summations run from 1 to N .

To solve for m and k we write the preceding equations in the following form:

$$(4) \quad \begin{cases} m \sum x_i + kN = \sum y_i \\ m \sum x_i^2 + k \sum x_i = \sum x_i y_i \end{cases}$$

By determinants,

$$(5) \quad \begin{cases} m = \frac{\begin{vmatrix} \sum y & N \\ \sum xy & \sum x \end{vmatrix}}{\begin{vmatrix} \sum x & N \\ \sum x^2 & \sum x \end{vmatrix}} = \frac{(\sum y)(\sum x) - N \sum xy}{(\sum x)^2 - N \sum x^2} \\ k = \frac{\begin{vmatrix} \sum x & \sum y \\ \sum x^2 & \sum xy \end{vmatrix}}{D} = \frac{(\sum x)(\sum xy) - \sum y \sum x^2}{D} \end{cases}$$

The determinant D in the expression for k is the same as that in the denominator of the expression for m . [In order to solve equations (4) for the values (5) it is assumed that D does not vanish.] The terms in the expressions for m and k refer to the original data. When these expressions have been evaluated they replace m and k in the equation $y = mx + k$.

Example 4. Find by the method of moments the best fitting line for the data in Example 2.

x	y	xy	x^2
0	66.7	0	0
1	72.7	72.7	1
2	82.3	164.6	4
3	92.1	276.3	9
4	93.0	372.0	16
5	100.6	503.0	25
15	507.4	1388.6	55

$$m = \frac{(507.4)(15) - 6(1388.6)}{(225) - 6(55)} = 6.86$$

$$k = \frac{15(1388.6) - 55(507.4)}{D} = 67.4.$$

Therefore,

$$y = 6.86x + 67.4.$$

7. An Alternative Procedure. In practice, it is sometimes easier to remember the procedure of fitting a line by the method of moments if one obtains the equations in (4) directly from the data instead of using the formulas for m and k . This will involve the following three steps:

(a) Substitute each of the given pairs of values in $y = mx + k$ and add the corresponding members of the resulting "equations." This gives the first equation in (4).

(b) Multiply each "equation" in (a) by the coefficient of m in that "equation" and add the corresponding members of the resulting "equations." This gives the second equation in (4).

(c) Solve the equations simultaneously. This will give the required values of m and k .

The algebraic statements which we designated "equations" (denoting that the statements are only approximately true) are called *observation equations* in the theory of errors. A linear combination of a set of linear observation equations is a true equation.

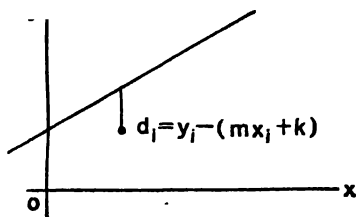
Example. Verify, for the data in Example 2, that the above procedure gives the same values of m and k as the formulas.

Step (a)	Step (b)
$66.7 = 0m + k$	$72.7 = m + k$
$72.7 = 1m + k$	$164.6 = 4m + 2k$
$82.3 = 2m + k$	$276.3 = 9m + 3k$
$92.1 = 3m + k$	$372.0 = 16m + 4k$
$93.1 = 4m + k$	$503.0 = 25m + 5k$
$100.6 = 5m + k$	
$507.4 = 15m + 6k$	$1388.6 = 55m + 15k$

Step (c)

Solving the equations, we obtain $m = 6.86$, $k = 67.4$, as before.

8. Least Squares. Case I. A standard method of fitting a curve to empirical data is one known as the method of *least squares*. As-



CASE I

sume, as before, that the plotted data suggest the linear relationship $y = mx + k$. Let d represent the difference between the ordinate of any given point and the corresponding ordinate of the line, that is, $d_i = [y_i - (mx_i + k)]$. These differences are called *residuals*. The

method of least squares is based upon the following principle.

PRINCIPLE OF LEAST SQUARES. The "best" estimate of a parameter is that for which the sum of weighted squares of the residuals is a minimum.¹

The sum is to be taken over all the observations that are subject to error. We shall assume that the observations are all of equal weight; consequently we may let each of the weights be unity. Then the parameters m and k are estimated by imposing the condition that $\sum_1^N d_i^2$ be a minimum. Now

$$(6) \quad \begin{aligned} \sum d^2 &= \sum [y - (mx + k)]^2 \\ &= Nk^2 + 2mk \sum x + m^2 \sum x^2 - 2k \sum y - 2m \sum xy + \sum y^2. \end{aligned}$$

This is a quadratic polynomial in k . We may write it in the form

$$(6a) \quad f(k) = Nk^2 + 2k(m \sum x - \sum y) + C$$

where C represents the terms not involving k . Then according to Theorem I the minimum value of $f(k)$ occurs when

$$k = \frac{\sum y - m \sum x}{N},$$

that is, when

$$Nk + m \sum x - \sum y = 0.$$

The right member of (6) is also a quadratic polynomial in m . We must choose m so that

$$m \sum x^2 + k \sum x - \sum xy = 0.$$

These last two equations² are the same as (4). When obtained by the method of least squares they are called *normal equations*. Therefore the values of m and k in (5) determine the best fitting line by both the method of moments and of least squares. It can be shown that the two methods give the same result for any polynomial.³

It is interesting to observe that the sum of the residuals is zero. Thus it can easily be shown that $\sum [y - (mx + k)] = 0$, when the

¹ For further information about this principle and a discussion of weights, the following books are recommended: (a) Reference 4. (b) *Statistical Mathematics* — A. C. Aitken. Oliver and Boyd.

² The student of calculus would obtain these equations as follows. Let $f(m, k) = \sum (y - mx - k)^2$. Then differentiate $f(m, k)$ partially with respect to m and k , respectively, and equate the results to zero.

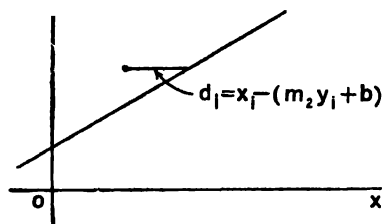
³ See *American Mathematical Monthly*, September, 1923.

values given in (5) are substituted for m and k . This property and the fact that the sum of the squares of the residuals is a minimum are quite analogous to two similar properties of the arithmetic mean, *viz.*,

(1) The sum of deviations from the mean is zero.

(2) The sum of the squares of deviations from the mean is less than the sum of the squares of such deviations taken from any other value, *i.e.*, $\mu_2 < \nu_2$.

Case II. In Case I distances between the points and the line were taken parallel to the y -axis. But we may just as logically, from a



CASE II

formal point of view, take distances parallel to the x -axis, and make the x residuals the basis for a least squares criterion of best fit. Similarly, for the method of moments: we can set up two equations such that the first moment of the observed x 's equals the first moment of the computed

x 's, and the zeroth moment of the observed x 's equals the zeroth moment of the computed. To do this let $x = m_2 y + b$ represent the equation of the line. Then by the principle of moments we have

$$\begin{aligned}\sum x &= \sum (m_2 y + b) \\ \sum xy &= \sum y(m_2 y + b).\end{aligned}$$

Solving for m_2 and b we obtain

$$(7) \quad \begin{cases} m_2 = \frac{\sum x \sum y - N \sum xy}{D} \\ b = \frac{\sum y \sum xy - \sum x \sum y^2}{D} \\ D = (\sum y)^2 - N \sum y^2. \end{cases}$$

If we determined m_2 and b by making the sum of the squares of the x residuals a minimum we would get the results given in (7). The expressions in (7) are those of (5) with x and y interchanged.

In general, Cases I and II will give different lines. Case I assumes that the observed points fail to fall on the line because of errors in the ordinates only. Case II assumes that only the x -coördinates are in error. In the application of curve fitting to economic data, etc., the formal mathematical procedure should not be used without

first verifying that the underlying assumptions involved in the procedure are justified. Inasmuch as the independent variable x can be controlled in experimental and observational data, the errors usually exist only in the y 's. Therefore, in speaking of the best line by the method of moments or least squares it is conventional to mean the line which fits best in the sense of (5) rather than (7).

Case III (for calculus students).

A third line can be obtained which fits best in the sense that the sum of the squares of the perpendicular distances from the points to the line is a minimum.

Let us suppose the equation of this line to be in the form

$$y' = mx' + k$$

where $x' = x - \bar{x}$, $y' = y - \bar{y}$, and (\bar{x}, \bar{y}) is the mean of the observed data. The distance d_i from this line to a point (x'_i, y'_i) representing a pair of observed values (referred to their respective means as origin) is, from analytics,

$$d_i = \frac{y'_i - mx'_i - k}{\sqrt{m^2 + 1}}.$$

We wish to make $\frac{1}{N} \sum_1^N d_i^2$ a minimum. Therefore we are to choose m and k so that the function

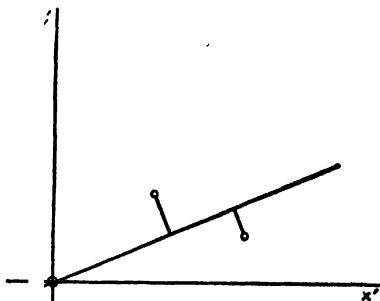
$$f(m, k) = \frac{1}{m^2 + 1} \left\{ \frac{1}{N} \sum_1^N (y'_i - mx'_i - k)^2 \right\}$$

is a minimum. This function may be written in the form

$$f(m, k) = \frac{1}{m^2 + 1} (\sigma_{y'}^2 + k^2 + m^2 \sigma_{x'}^2 - 2mr\sigma_{y'}\sigma_{x'})$$

where r is a convenient symbol defined by the relation

$$r\sigma_{y'}\sigma_{x'} = \frac{1}{N} \sum_1^N x'_i y'_i.$$



CASE III

To make $f(m, k)$ a minimum we first put $k^2 = 0$. Then we equate to zero the first derivative with respect to m and obtain

$$m^2 r \sigma_y \sigma_x - m(\sigma_y^2 - \sigma_x^2) - r \sigma_y \sigma_x = 0.$$

Solving for m we have

$$m = \frac{(\sigma_y^2 - \sigma_x^2) \pm [(\sigma_y^2 - \sigma_x^2)^2 + 4r^2 \sigma_y^2 \sigma_x^2]^{1/2}}{2r \sigma_y \sigma_x}$$

Therefore the required equation is $y' = mx'$. Referred to the origin of x and y , this is

$$y - \bar{y} = m(x - \bar{x})$$

where m is determined above.

This line is the appropriate one to fit if there are errors in both x and y of the empirical data.

A special problem under Case I. Sometimes problems arise where the line to be fitted is restricted in some way. For example, the nature of the problem may require that the line shall pass through the origin. If this condition is imposed, (1) takes the form

$$y = mx.$$

The least squares estimate of the slope of this line depends upon various assumptions about the errors. If y is subject to error and x is free of error, and if the observations are all of equal weight, it is easy to show that

$$m = \frac{\sum xy}{\sum x^2}$$

by the principle of least squares. This principle will give different estimates of m under different assumptions about the weights of the observations. Several particular solutions of the more general problem and some applications will be found in §15 of reference 4 on page 6. (See also our Exercise 11, p. 189.)

Exercises

1. Fit a line to the following data by Case I:

Ans. $y = -.5x + 8$.

x	6	7	7	8	8	8	9	9	10
y	5	5	4	5	4	3	4	3	3

2. Show that $\sum d = 0$ for Exercise 1.
3. Using the values given in (5) for m and k show that $\sum [y - (mx + k)] = 0$.
4. Verify the expressions for m_2 and b given in (7). How would you modify the "alternate procedure" so it will apply to m_2 and b ?
5. Fit a line to the data of Example 2 by the method of Case II.
6. Show that the formulas in (5) fail when the x 's are all equal. *Hint.* Replace x by a constant c in the denominator D .

9. Simplification. The formulas for m and k may be simplified. For certain purposes it may be desirable to make the transformations $x' = x - \bar{x}$ and $y' = y - \bar{y}$. This has the effect, graphically, of

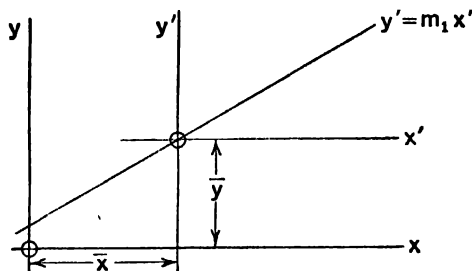


FIG. 27

translating the origin to the point (\bar{x}, \bar{y}) so that the y -axis is moved to the value \bar{x} , and the x -axis is moved to the value \bar{y} . Let the equation of the line with reference to these new axes be $y' = m_1 x' + k_1$. The formulas for m_1 and k_1 will be the same as for m and k except that x will be replaced by x' and y by y' . Hence

$$m_1 = \frac{N \sum x'y' - \sum x' \sum y'}{N \sum x'^2 - (\sum x')^2}$$

$$k_1 = \frac{\sum x'^2 \sum y' - \sum x' \sum x'y'}{N \sum x'^2 - (\sum x')^2}.$$

But since x' is a deviation from the mean of x , $\sum x' = 0$. Similarly, $\sum y' = 0$. Hence the values of m_1 and k_1 reduce to

$$(8) \quad m_1 = \frac{\sum x'y'}{\sum x'^2}, \quad k_1 = 0.$$

Therefore the line goes through the new origin, and its equation is

$$(9) \quad y' = m_1 x'$$

where m_1 is defined in (8).

The above transformation may not lighten the computations unless the values of x or y are equispaced. However, it does simplify the theory in certain applications, particularly in correlation theory (Chapter VIII).

10. Time Series. If one of the variables is time, as in Examples 1 and 2, the data are called a time series. The best fitting line is then commonly called a trend line or trend. In the process of fitting a trend line, a first simplification, obviously, is to take the origin at one of the given dates as we did in Example 3. But a much greater simplification is possible, if the x 's are equispaced, as they usually are in a time series. Denote the common differences of the x 's by c and the mid-date by \bar{x} . Then we may shift the origin to \bar{x} and change the unit of measurement along the horizontal axis to c . Thus we may let

$$(10) \quad t = \frac{x - \bar{x}}{c}$$

where

$$(11) \quad \bar{x} = \frac{x_1 + x_N}{2}$$

if the x 's are equispaced.

Let us think now of our line in (t, y) coördinates, and let its equation be $y = at + b$. Our problem is to find a and b numerically from the given data, as we found m and k before. Our normal equations will be

$$\begin{aligned} \sum y &= \sum (at + b) \\ \sum ty &= \sum (at + b)t. \end{aligned}$$

Since $\sum t = \frac{1}{c} \sum (x - \bar{x}) = 0$, and $\sum b = Nb$, the above equations are readily solved, giving

$$(12) \quad a = \frac{\sum ty}{\sum t^2}, \quad b = \frac{1}{N} \sum y.$$

The student should remember that this simplification can be used only when the x 's are equispaced.

Example 5. Find the trend line for the following data. Here $c = 5$, and from (11) $\bar{x} = 10$.

x	y	t	ty	t^2
0	12	-2	-24	4
5	15	-1	-15	1
10	17	0	0	0
15	22	1	22	1
20	24	2	48	4
Sums	90		31	10

From (12),

$$a = \frac{31}{10} = 3.1, \quad b = \frac{90}{5} = 18.$$

So the required equation is $y = 3.1t + 18$, with reference to the new origin and units. If we wish it in terms of x , we substitute

$$t = \frac{x - 10}{5}$$

and obtain $y = .62x + 11.8$.

Example 6. Same as Example 5, with another observation added. Note that when there is an even number of observations, the values of t are fractional. In this case it is convenient to use the column headings $2ty$ instead of ty , and $4t^2$ instead of t^2 .

x	y	t	$2ty$	$4t^2$
0	12	-5/2	-60	25
5	15	-3/2	-45	9
10	17	-1/2	-17	1
15	22	1/2	22	1
20	24	3/2	72	9
25	30	5/2	150	25
Sums	120		122	70

$$\bar{x} = 12.5, \quad \sum ty = 61, \quad \sum t^2 = 17.5$$

$$a = 3.49, \quad b = 20$$

$$y = 3.49t + 20$$

$$y = 3.49 \left(\frac{x - 12.5}{5} \right) + 20$$

$$y = .7x + 11.28.$$

11. Exponential Trends. When the given y values form a geometric progression while the corresponding x values form an arithmetic progression, the relationship between the variables is given by an exponential function, and the best fitting curve is said to describe an exponential trend. Data from the fields of biology, banking, and economics frequently exhibit such a trend. Thus the growth of bacteria is exponential. Money accumulating at compound interest follows the same kind of law of growth. And in business, sales or earnings may grow exponentially over a short period. Another familiar example is the increase in friction as a rope is coiled around a post. As the number of coils increases in arithmetic progression, the friction increases in geometric progression.¹ This explains why a few turns of the hawsers around the bitts at the wharf is sufficient to hold a large ship.

The characteristic property of this law is that the rate of growth, that is, the rate of change of y with respect to x , at any value of x is proportional to the value of the function for that value of x . The function

$$(13) \quad y = Ae^{Bx}$$

has this property.² The letter e is a fixed constant, whereas A and B are parameters to be determined from the data. If y decreases as x increases, B is negative. An interesting example of this case is the disappearance of radioactive substances like radium.

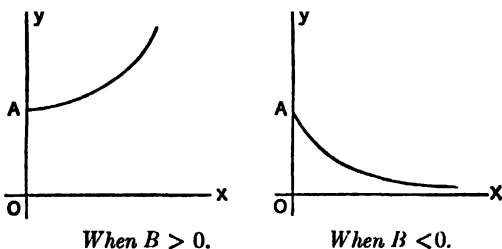


FIG. 28 — GENERAL APPEARANCE OF THE GRAPH OF (13) FOR $x \geq 0$ AND $A > 0$.

To assume that the apparent law of growth will continue is usually unwarranted, so only short range predictions can be made with any considerable degree of reliability. When the exponential character

¹ *Elementary Mathematical Analysis* — C. S. Slichter. McGraw-Hill.

² The student of calculus will understand that "rate of change" is used here in the derivative sense. For (13), $dy/dx = ky$.

of the observed phenomenon ceases a saturation point is said to be reached.

The parameters A and B. If we transform (13) so that it is linear with respect to its parameters we may use the methods for fitting a straight line to determine A and B . To this end we first take the logarithms (to base 10) of both sides of (13), obtaining

$$(14) \quad \log y = \log A + (B \log e)x$$

which is of the form

$$(15) \quad Y = k + mx$$

where $Y = \log y$, $k = \log A$, $m = B \log e$.

If we look up the logarithms of the given y 's and denote them by Y , we may fit the equation $Y = mx + k$ to the (x, Y) values by determining m and k by means of the formulas given in (5). In using these formulas we must remember to replace y by Y . After m and k are determined, A and B may be obtained from the relations

$$A = \text{anti-log of } k$$

$$B = m / \log e, \text{ where } \log e = \log 2.718 \\ = .4343.$$

The student may be interested to verify that the relation $Y = mx + k$ can be put back into the form (13). We may write (14) in the form

$$y = 10^{\log A + (B \log e)x} \\ = \{10^{\log A}\} \{10^{\log e}\}^{Bx} \\ = Ae^{Bx}.$$

The last step follows because $10^{\log_{10} N} = N$ by definition of logarithm.

Example 7. Find the exponential trend for the following data, and draw the curve.

x	y	Y	xY	x^2
1	1.6	.2041	.2041	1
2	4.5	.6532	1.3064	4
3	13.8	1.1399	3.4197	9
4	40.2	1.6042	6.4168	16
5	125.0	2.0969	10.4845	25
15		5.6983	21.8315	55

From (5) we have,

$$D = (\sum x)^2 - N \sum x^2$$

$$m = \frac{1}{D} [\sum Y \sum x - N \sum xY]$$

$$k = \frac{1}{D} [\sum x \sum xY - \sum Y \sum x^2].$$

Therefore,

$$D = [(15)^2 - 5(55)] = -50$$

$$m = \frac{1}{D} [(5.6983)(15) - 5(21.8315)] = .4737$$

$$\begin{aligned} k &= \frac{1}{D} [15(21.8315) - (5.6983)(55)] \\ &= -.2813 = 9.7187 - 10. \end{aligned}$$

And

$$\log A = 9.7187 - 10, \text{ hence } A = .5232$$

$$B = \frac{m}{.4343} = 1.091.$$

Therefore the required equation is

$$y = .5232e^{1.091x}.$$

When the x 's are equispaced, as here, the work may be simplified by using (10) and fitting a line

$$Y = at + b.$$

The problem now is essentially the same¹ as in §10 where a and b are defined in (12) except that we are now dealing with (t, Y) coördinates instead of (t, y) .

The method is illustrated below.

t	Y	tY	t^2
-2	.2041	-.4082	4
-1	.6532	-.6532	1
0	1.1399	0.0000	0
1	1.6042	1.6042	1
2	2.0969	4.1938	4
$t = x - 3$	5.6983	4.7366	10

¹ The critical reader will realize that fitting a straight line to the values of $\log y$ is not quite the same as fitting an exponential to the values of y . However, the discrepancy usually does not affect the fit seriously. For a method which is free from this difficulty, see *Glover's Tables*, p. 468.

From (12)

$$a = \frac{\sum tY}{\sum t^2} = \frac{4.7366}{10} = .4737$$

$$b = \frac{1}{N} \sum Y = \frac{5.6983}{5} = 1.1397.$$

So

$$Y = 4737t + 1.1397.$$

Transforming this into (x, Y) coördinates we have

$$Y = .4737(x - 3) + 1.1397$$

$$= .4737x - .2814$$

as before.

For purposes of plotting, predicting, or interpolating, values of y in (13) may be obtained by means of the intermediate form (15). So, to sketch the curve

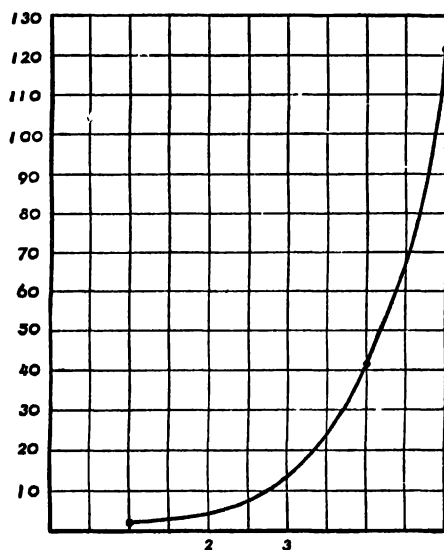


FIG 29

for this example, we first assign values to x in the last equation, compute the corresponding values of Y , and then obtain the values of y from a table of logarithms. These values are given in the following table. The curve in Figure 29 is sketched from the (x, y) values in this table.

x	1	2	3	4	5	6
Y	0.1923	0.6660	1.1397	1.6134	2.0871	2.5608
y	1.56	4.63	13.79	41.06	122.2	363.8

12. Further Remarks on the Exponential Function. Equation (13) is sometimes called the compound interest law because it describes the way money would grow if interest were compounded continuously. If P dollars are invested at a nominal rate $j\%$ compounded m times a year, the amount S after x years is given by the formula

$$S = P \left(1 + \frac{j}{m} \right)^{mx}.$$

If j is compounded continuously or, in other words, if m is taken indefinitely large (written $m \rightarrow \infty$), the amount S does not increase indefinitely but approaches a limiting value. We may write the expression for S in the form

$$S = P \left[\left(1 + \frac{j}{m} \right)^{m/j} \right]^{jx}.$$

If we let $N = m/j$, we have

$$S = P \left[\left(1 + \frac{1}{N} \right)^N \right]^{jx}.$$

It can be shown in the calculus¹ that, as $N \rightarrow \infty$, the quantity

$\left(1 + \frac{1}{N} \right)^N$ approaches the limit called e . Thus we have

$$\lim_{N \rightarrow \infty} \left(1 + \frac{1}{N} \right)^N = e = 2.718$$

This limit is also the base of the Napierian, or natural, system of logarithms. As $m \rightarrow \infty$ so does $N \rightarrow \infty$. Therefore in the ideal case of continuous conversion of interest, we have the limiting form

$$\begin{aligned} S &= \lim_{m \rightarrow \infty} P \left[\left(1 + \frac{j}{m} \right)^{m/j} \right]^{jx} \\ &= \lim_{N \rightarrow \infty} P \left[\left(1 + \frac{1}{N} \right)^N \right]^{jx}, \end{aligned}$$

that is

$$S = Pe^{jx}$$

which is of the form (13).

There are several other forms of the exponential function. For example, if we let $r = e^B$, (13) becomes

$$y = Ar^x$$

¹ The teacher can give appropriate references.

which is the general term of a geometric progression whose first term is A and common ratio is r .

If B is negative in $r = e^B$ then $r < 1$. So (13) is a decreasing function when B is negative.

If we let $10^k = e^B$, (13) becomes

$$y = A10^{kx}.$$

Then $k = B \log_{10} e$ and k differs from B by the factor $\log_{10} e$. This factor is known as the *modulus* of the system of logarithms of base 10 with respect to the system of base e .

The value of the reciprocal of the modulus

$$\frac{1}{\log_{10} e} = 2.3025851 \dots$$

is often useful. For example, suppose that the logarithm to base e is required for a given number N and tables to base 10 only are available. Let $\log_e N = x$. Then $e^x = N$, and $x \log_{10} e = \log_{10} N$, whence $x = \log_{10} N / \log_{10} e = 2.303 \log_{10} N$. (Hereafter, the base 10 will be understood unless otherwise indicated.)

13. Ratio Charts. In the graphical representation of data that exhibit an exponential trend, it is often desirable to use semi-logarithmic paper. Such paper has a logarithmic scale in the vertical direction and a uniform scale in the horizontal direction. (Figure 30.) A logarithmic scale is one in which the distance from $y = 1$ to $y = N$ equals $\log N$. A "cycle" of rulings spaced according to the logarithms of the integers from 1 to 10 is the unit of the vertical $\log y$ scale.

"Semi-log" paper may be constructed or purchased having one or more cycles. The appropriate number of cycles is determined by the range of y values in the data to be plotted. If the bottom line of the first cycle is labeled 1 and taken as the origin of $\log y$ ($\log 1 = 0$), the beginning of the next cycle is read 10 ($\log 10 = 1$), the next one above that is read 100 ($\log 100 = 2$), etc. However, the beginning of the first cycle may be labeled with any number which is an integral power (positive or negative) of 10, as .01, .1, 10, 100, etc. Corresponding lines in successive cycles are labeled with numbers which are 10 times those in the preceding cycle. Since y has no real logarithm if $y \leq 0$, neither zero nor negative numbers are found on a logarithmic scale. Plotting a point whose semi-logarithmic co-

ordinates are (x, y) is equivalent to plotting the point whose rectangular coördinates are $(x, \log y)$.

Example 8. Plot $y = 8 (2^x)$ on semi-log paper.

Solution. Assigning values to x we form the following table,

x	-3	-2	-1	0	1	2	3	4
y	1	2	4	8	16	32	64	128

from which we obtain the semi-logarithmic graph shown in Figure 30.

We now state the following theorem.

Theorem II. *If A is a positive constant, the $(x, \log y)$ -graph of $y = Ae^{Bx}$ is a straight line.*

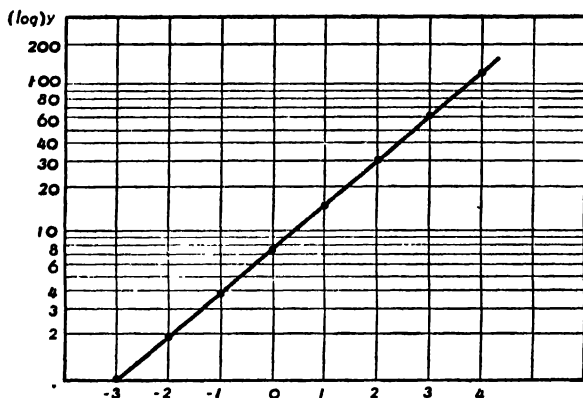


FIG. 30

Proof: Since (15) is linear in x and Y , its graph in (x, Y) rectangular coördinates is a straight line.

Semi-logarithmic graphs are also called *ratio charts*. Their usefulness depends upon the property of logarithms that

$$\log \frac{M}{N} = \log M - \log N.$$

It follows that the distance between any two ordinates of the chart measures the ratio between the values represented by these ordinates.

Thus if

$$\frac{y_1}{y_2} = \frac{y_3}{y_4}$$

then

$$\log y_1 - \log y_2 = \log y_3 - \log y_4$$

or

$$Y_1 - Y_2 = Y_3 - Y_4,$$

that is, equal ratios are represented by equal vertical distances. Likewise, if

$$\frac{y_1}{y_2} > \frac{y_3}{y_4}$$

then

$$Y_1 - Y_2 > Y_3 - Y_4$$

and the larger ratio is represented graphically by the larger distance. These differences of elevation are independent of any base line. The same percentage increase in y is represented by the same addition to the height of Y in all parts of the chart. Hence, it is easier to depict and discover percentage changes on ratio charts than on ordinary charts.

The analysis of time series in economic statistics is often facilitated by forming "link relatives" which are ratios of each ordinate (after the first) to the preceding ordinate. Thus, if y_1, y_2, \dots, y_n are the given values, the link relatives are

$$R_1 = \frac{y_2}{y_1}, \quad R_2 = \frac{y_3}{y_2}, \quad \dots, \quad R_{n-1} = \frac{y_n}{y_{n-1}}.$$

Any link relative R denotes the percentage change in y from one month (say) to the next. If the y 's are plotted on ratio paper they will lie on a straight line when the R 's are equal, on a curve bending upward when the R 's are increasing, and on a curve bending downward when the R 's are decreasing. It follows that if two curves are parallel on ratio paper their rate of increase (or decrease) is the same.

For further discussion of ratio charts the student is referred to the books of Bivins and Haskell (see §7, Introduction).

Graphical determination of exponential function. It follows from Theorem II that data giving a straight line when plotted on semi-logarithmic paper (with x on the uniform and y on the logarithmic scale) satisfy an equation of the form (13). Suppose that the

(straight line) graph has been drawn and one desires the exponential function which the line represents and the data satisfy. The constants A and B in (13) can be approximated by the following method.¹ We first observe that the slope of the line represented by (15) is given by

$$m = B \log e = \frac{Y_2 - Y_1}{x_2 - x_1}.$$

To determine the numerical value of B , take one cycle of y (over which the graph extends) from any starting point and read the corresponding values of x (Figure 31a), so that

$$B = \frac{Y_2 - Y_1}{(x_2 - x_1) \log e} = \frac{\log (y_2/y_1)}{\Delta x \log e} = \frac{\log 10}{\Delta x \log e} = \frac{2.303}{\Delta x}.$$

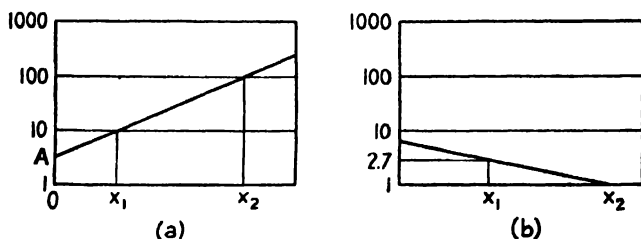


FIG. 31

In case the graph does not extend over one cycle, determine x for $y = e$ and $y = 1$; then (Figure 31b)

$$B = \frac{\log e}{\Delta x \log e} = \frac{1}{\Delta x}.$$

The sign of B is of course positive if the graph has a positive slope in the ordinary sense and is negative for a negative slope.

If the graph intersects the line $x = 0$, the value of A can be read off at this intersection. If, in the data involved, the graph does not intersect the line $x = 0$, A can usually be determined by finding y for some convenient values of x such as $Bx = \text{some integer } n$, whereupon $A = y/e^n$ from equation (13).

In practical problems, the plotted points representing the data

¹ Note on *Semi-Logarithmic Graphs*—W. T. Lenser, *The American Mathematical Monthly*, vol. 49 (1942), pp. 611–613.

will not usually fall exactly on a straight line. But if they exhibit a linear trend one may draw (with the aid of a transparent ruler) the line that seems to fit them best. Then proceed as above.

Example 9. The uniform scale along the horizontal axis of a sheet of semi-logarithmic paper ranges from 0 to 10; along the vertical axis the logarithmic scale ranges from 100 to 1000. A straight line is drawn on the paper from the upper endpoint of the vertical scale to the midpoint of the horizontal scale. Determine (i) the equation of the exponential function represented by the line, (ii) the equation of the line in (x, Y) coördinates.

Solution 1, using above method. $A = 1000$. $B = -1/(5 \log e) = (-2.3)/5 = -0.46$. Hence, the desired equation (i) is

$$y = 1000e^{-0.46x}.$$

The slope of the line is $m = B \log e = -\frac{1}{5}$ and its equation (ii) is

$$Y = 3 - 0.2x.$$

Solution 2. The line goes through the points (0, 1000) and (5, 100). Substitution of the first pair of coördinates into (13) gives $A = 1000$. Substitution of the second pair into $y = 1000e^{Bx}$ gives $100 = 1000e^{5B}$. Then $e^{-5B} = 10$ and $-5B = \log_e 10 = 2.303$, whence $B = -0.46$.

14. Logarithmic Coördinate Paper. A function of the form

$$(16) \quad y = kx^m$$

is called a *power function*. If $k > 0$ we have

$$(17) \quad Y = K + mX$$

where the capital letters denote the logarithms of the corresponding lower-case letters. Form (17) suggests the usefulness of *logarithmic coördinate* paper on which the rulings in both directions are at distances from the origin that are proportional to the logarithms of the numbers represented. To mark on this paper a point whose ordinary coördinates are (X_1, Y_1) we plot the point whose rulings correspond to the numbers x_1 and y_1 .

It is evident from (17) that the graph of (16) is a straight line on logarithmic coördinate paper. It also follows from (17) that the problem of fitting a curve of the form (16) to a set of observations can be reduced to the problem of fitting a straight line.

Example 10. A straight line is drawn on logarithmic coördinate paper through the points (4, 16) and (6, 54). Determine the function $y = f(x)$ which has that line as its graph.

Solution 1. Substitution of the coördinates of the given points into (16) gives

$$\begin{cases} 16 = k(4^m) \\ 54 = k(6^m) \end{cases}$$

Upon dividing each member of the first equation by the corresponding member of the second, we obtain $8/27 = (2/3)^m$ whence by inspection $m = 3$. Then $k = \frac{1}{4}$, and the required function is $4y = x^3$.

Solution 2. Substitution of the logarithms of the given coördinates into (17) gives

$$\begin{cases} 1.20412 = K + 0.60206m \\ 1.73239 = K + 0.77815m \end{cases}$$

Solving, $m = 3$ and $K = -.60206 = 9.39794 - 10$, $k = .25$.

15. Parabolic Trend. Data of broad economic or social significance extending over a long period of years may often be described by an arc of a second degree parabola. The equation of a parabola is of the form

$$y = \alpha + \beta x + \gamma x^2$$

where α, β, γ are the parameters to be determined.

If the x 's are equispaced we may let

$$t = \frac{x - \bar{x}}{c},$$

where $\bar{x} = (x_1 + x_N)/2$ and $c = |x_{i+1} - x_i|$, and thereby effect considerable simplification in evaluating the constants. In t and y coördinates the equation will, of course, involve different constants and we may write its equation in the form

$$(18) \quad y = A + Bt + Ct^2.$$

The method of moments may again be used and since (18) is a polynomial this method also gives the best fitting curve in a least squares sense. Because there are three constants to be determined we must equate the second moments as well as the zeroth and first moments. Imposing these conditions of moments between the observed and computed ordinates, we obtain the three normal equations:

$$\begin{aligned} \sum y &= NA + B\sum t + C\sum t^2 \\ \sum ty &= A\sum t + B\sum t^2 + C\sum t^3 \\ \sum t^2y &= A\sum t^2 + B\sum t^3 + C\sum t^4. \end{aligned}$$

Since the mean is chosen as origin $\sum t = 0$. With this choice of origin and because the x 's are equispaced it can be shown that $\sum t^3 = 0$. Therefore the normal equations simplify into

$$(19) \quad \begin{cases} B = \frac{\sum ty}{\sum t^2} \\ AN + C\sum t^2 = \sum y \\ A\sum t^2 + C\sum t^4 = \sum t^2 y. \end{cases}$$

When the summations involved in these equations are evaluated from the data the values of A , B , and C can easily be determined.

Example 11. Fit a parabola to the following data.

NUMBER OF DIVORCES PER 1000 MARRIAGES IN THE UNITED STATES
1900-1930

Year	y	x	t	ty	t^2	t^2y	t^4
1900	81	0	-3	-243	9	729	81
1905	84	5	-2	-168	4	336	16
1910	88	10	-1	-88	1	88	1
1915	104	15	0	0	0	0	0
1920	134	20	1	134	1	134	1
1925	148	25	2	296	4	592	16
1930	170	30	3	510	9	1530	81
Sums	809	$\bar{x} = 15$		441	28	3409	196

From (19),

$$B = \frac{441}{28}$$

$$7A + 28C = 809$$

$$28A + 196C = 3409.$$

Solving the last two equations simultaneously we obtain,

$$A = \frac{322}{3}, \quad C = \frac{173}{84}.$$

Therefore,

$$y = \frac{322}{3} + \frac{441}{28}t + \frac{173}{84}t^2.$$

If we desire the equation in the original form we substitute $t = \frac{1}{3}(x - 15)$ and obtain

$$y = \frac{322}{3} + \frac{441}{28} \left(\frac{x - 15}{5} \right) + \frac{173}{84} \left(\frac{x - 15}{5} \right)^2$$

which simplifies into

$$y = 78.62 + .68x + .0824x^2.$$

Upon the hypothesis that divorces will continue to increase according to this trend, we may estimate the number for 1950 for example. When $x = 50$ in the above equation, we find $y = 318.62$.

16. The Gompertz Curve. The curve which bears his name was suggested in 1825 by Gompertz for use in actuarial science. Recently it has had some application as a *growth curve* in business and population forecasting and in certain problems in education. Its equation¹ is

$$(20) \quad y = kg^{x^2}.$$

To determine the parameters, we first transform (20) into the logarithmic form

$$(20a) \quad Y = K + Gc^x$$

where $Y = \log y$, $K = \log k$, $G = \log g$. The number, N , of observations available must be such that $N = 3n$ where n is the number in each of three subgroups with no observations omitted; that is, N must be divided into three blocks of data consisting of n items each. It is also necessary that the values of the independent variable x be equispaced. Then the origin can be chosen so that x takes the values $0, 1, 2, \dots, 3n - 1$. If these values of x are substituted in (20a) we obtain the three sets of functional Y 's shown in (a), (b), and (c).

$$\left. \begin{array}{cc} 0 & Y_0 \\ 1 & Y_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ n-1 & Y_{n-1} \end{array} \right\} \sum_{i=0}^{n-1} Y_i, \quad \left. \begin{array}{l} Y_0 = K + Gc^0 \\ Y_1 = K + Gc \\ \cdot \\ \cdot \\ Y_{n-1} = K + Gc^{n-1} \end{array} \right\} \quad (a)$$

$$\left. \begin{array}{cc} n & Y_n \\ n+1 & Y_{n+1} \\ \cdot & \cdot \\ \cdot & \cdot \\ 2n-1 & Y_{2n-1} \end{array} \right\} \sum_{i=n}^{2n-1} Y_i, \quad \left. \begin{array}{l} Y_n = K + Gc^n \\ Y_{n+1} = K + Gc^{n+1} \\ \cdot \\ \cdot \\ Y_{2n-1} = K + Gc^{2n-1} \end{array} \right\} \quad (b)$$

$$\left. \begin{array}{cc} 2n & Y_{2n} \\ 2n+1 & Y_{2n+1} \\ \cdot & \cdot \\ \cdot & \cdot \\ 3n-1 & Y_{3n-1} \end{array} \right\} \sum_{i=2n}^{3n-1} Y_i, \quad \left. \begin{array}{l} Y_{2n} = K + Gc^{2n} \\ Y_{2n+1} = K + Gc^{2n+1} \\ \cdot \\ \cdot \\ Y_{3n-1} = K + Gc^{3n-1} \end{array} \right\} \quad (c)$$

¹ For a derivation see *Mathematical Theory of Life Insurance* — Forsyth. John Wiley and Sons, Inc.

Let S_1, S_2, S_3 denote respectively the totals of the subgroups (a), (b), and (c). Thus we have

$$\begin{aligned} S_1 &= nK + G(1 + c + \cdots + c^{n-1}) \\ S_2 &= nK + Gc^n(1 + c + \cdots + c^{n-1}) \\ S_3 &= nK + Gc^{2n}(1 + c + \cdots + c^{n-1}). \end{aligned}$$

Then

$$\begin{aligned} S_2 - S_1 &= G(c^n - 1)(1 + c + \cdots + c^{n-1}) \\ S_3 - S_2 &= Gc^n(c^n - 1)(1 + c + \cdots + c^{n-1}) \end{aligned}$$

whence we obtain

$$c^n = \frac{S_3 - S_2}{S_2 - S_1}.$$

Writing the expression for $S_2 - S_1$ in the form

$$S_2 - S_1 = G \frac{(c^n - 1)^2}{c - 1}$$

and solving for G , we obtain

$$G = \frac{(S_2 - S_1)(c - 1)}{(c^n - 1)^2}.$$

The expression for S_1 may be written

$$S_1 = nK + \frac{G(1 - c^n)}{1 - c},$$

so we have

$$K = \frac{1}{n} \left[S_1 - \frac{G(1 - c^n)}{1 - c} \right].$$

In the above expressions, S_1, S_2, S_3 denote sums of the functional Y 's. If these are now replaced by the empirical data so that

$$S_1 = \sum_n^{n-1} Y_i, \quad S_2 = \sum_n^{2n-1} Y_i, \quad S_3 = \sum_{2n}^{3n-1} Y_i,$$

where Y_i refers to the observed Y 's, then c can be determined from the expression for c^n . Using the value of c , G can be determined, and then K .

If $c < 1$, it is clear from (20a) that $Y \rightarrow K$ as $x \rightarrow \infty$. Then $y = k$ is an asymptote and k is sometimes called the *ceiling* of the curve. (See Figure 32.)

For an application of the above method to a problem in business, see *Statistical Methods (Revised Edition)* by Mills, page 672.

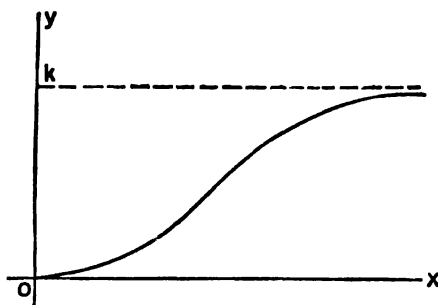


FIG. 32

17. Remarks and References. The methods of least squares and moments do not select the appropriate curve. They merely determine the "best" values of the parameters in the equation of the curve which has been selected previously to describe the observed data. The question of the type of curve which should be fitted to the data is not always easy to answer. The selection of the appropriate mathematical function depends to a large extent upon the investigator's experience in the field in which the problem lies and his knowledge of the properties of curves. It always helps to plot the data first. The usual requirements for practical purposes are that (a) the curve must represent well the trend of the empirical data, and (b) the mathematical expression must not involve too many parameters and those present must be calculable from the data. In dealing with time series, if the objective is to find out what would happen if the *percentage* change should continue as it has on the average in the past, then an exponential trend is indicated. If the objective is to find out what would happen if the yearly (or monthly, etc.) change should continue as it has in the past, a straight line trend is indicated.

We will merely mention here two other important curves which require more advanced mathematics in their treatment. The *logistic*, or so-called Reed-Pearl curve, is used extensively in studying various growth phenomena. Its function is of the form

$$y = \frac{1}{a + bc^x}$$

and it resembles somewhat the Gompertz curve discussed above. For further discussion of this curve and methods of fitting it see

1. *Elements of Statistics* — Davis and Nelson.
2. *Statistical Methods, Revised* — Mills.

The function

$$y = ks^x q^{c^x}$$

is known as Makeham's law. It is used in actuarial work. The student having a working knowledge of the calculus will find an interesting discussion of its use in the field of insurance in an article entitled *Makeham's Laws of Mortality*, Rietz, *American Mathematical Monthly*, vol. 28, p. 471.

The logistic curve was used in studies on the rate of growth of the population of the United States. But its usefulness in this connection fell somewhat short, apparently,¹ of the claims of its sponsors. Two other references relating to the population of our country may appropriately be mentioned here. Although they do not involve problems of curve fitting they do afford instructive examples of the application of scientific method to social and political problems. They are

1. *Bibliography on Methods of Apportionment in Congress* — E. V. Huntington. *American Mathematical Monthly*, vol. 49 (1942), pp. 115-117.

2. *Determination of the Center of Population in the United States*. *School Science and Mathematics*, May and June, 1942.

Exercises

1. If the rate of change of y with respect to x is always proportional to the attained value of y then y is what kind of a function of x ?
2. Determine A and B in the best fitting curve of the type (13) for the following data.

Data		Form for Computations			
x	y	t	Y	tY	t^2
0	1000				
5	100				
10	10				
15	1				
20	.1				

3. (a) Prove formula (11).
(b) Graph the curve $y = 10e^{-2x}$.
4. Find the best fitting parabola for the following points: $(-4, 2)$, $(0, 8)$, $(4, 9)$, $(8, 11)$, $(12, 8)$, $(16, 5)$. Ans. $y = 7.2 + .94x - .07x^2$.

¹ *Differential Equations Subject to Error, and Population Estimates* — Harold Hotelling. *Jour. Amer. Stat. Assoc.*, vol. 22 (1927), pp. 283-314.

5. If the values of t form an arithmetic progression and $\sum t = 0$ prove that $\sum t^2 = 0$.
6. (a) Add the values $x = 30, y = 37$ to the data of Example 6 and find the trend line. *Ans.* $y = .8x + 10.43$.
- (b) On the hypothesis that the apparent trend continues, predict the value of y when $x = 35$
7. In a tensile test of a metal bar the following observations were made, where x represents the load in tons and y the elongation in ten-thousandths of an inch:

x	1	2	3	4	5
y	14	27	40	55	68

Determine a linear relation between x and y by the theory of least squares.

8. In the following table y represents the fire losses in the United States in millions of dollars. Taking the origin of x at 1915 find the best fitting line, in a least squares sense, for the data.

x	1915	1917	1919	1921	1923	1925
y	172	290	321	495	535	570

9. (a) Add the values $x = 6, y = 300$ to the data of Example 7 (p. 153) and find the equation of the best fitting exponential curve.

$$\text{Ans. } Y = .4617x - .2534$$

$$y = .56e^{1.06x}.$$

- (b) Plot the given data and the curve obtained in (a) on semi-log paper.
10. Distinguish between the forms of the curves represented by the functions $y = Ae^{-Bx}$ and $y = Ke^{-hx^2}$ where A, B, K , and h are positive real numbers. If these functions were plotted on semi-log paper what kind of curves would be obtained?
11. Determine by inspection the value of (a) $10^{\log_{10} e}$, (b) $a^{\log_a N}$.
12. Solve for x : $\log_{10}(x^2) = (\log_{10} x)(\log_e x)$.
13. Solve for x : $\log_{10}(x^2) - \log_{10}(x/10) = 2$.
14. Determine a number x such that the square of $\log x$ exceeds $\log x$ by 2. (Logs to base 10. Two answers.)
15. On semi-logarithmic coordinate paper, a straight line is drawn through the points (2, 1) and (4, 100). Determine the function which has that line as its graph. *Hint.* Use the form $y = Ar^x$. *Ans.* $100y = 10^x$.
16. Same as exercise 15 for the points (1, 6) and (2, 18). *Ans.* $y = 2(3^x)$.

17. On logarithmic coordinate paper, a straight line is drawn through the points (2, 12) and (3, 27). Determine the function which has that line as its graph. *Ans.* $y = 3x^2$.
18. Data from a certain experiment involving voltage (v) as a function of time (t) are plotted on logarithmic coordinate paper, and are found to exhibit a linear trend there. A line is drawn, with a transparent ruler, which seems to fit the plotted data best. Two points on this line are (6, 18) and (8, 32). Determine an equation expressing v in terms of t whose logarithmic graph is the line.
19. Draw the graph of $y = 25x^n$ on logarithmic coordinate paper, (a) when $n = 2$, (b) when $n = -2$. Mark scales clearly.
20. The graph of $y = \log_{10} x$ assists one in remembering several important properties of the logarithms of real numbers. Sketch this graph and state some of these properties.
21. Read and report on one or more of the references cited in §17.

Note. Source material for additional exercises on curve fitting may be found in the current volumes of the following publications:

1. *Statistical Abstract of the United States.*
2. *World Almanac and Book of Facts.*

CHAPTER VIII

CORRELATION THEORY

1. The Meaning of Simple Correlation. So far we have been concerned with the problems which arise from variation in a single variable. We will now consider the simultaneous variation of two variables. Methods for disclosing the facts of co-variation and for measuring the degree of relationship existing between two variables are due mainly to the English biometricians Sir Francis Galton (1822-1911) and Karl Pearson (1857-1936).

Data presenting two sets of related measurements or observations may arise in many fields of activity yielding N pairs of corresponding variates (x_i, y_i) , $i = 1, 2, 3, \dots, N$. Thus x may represent July rainfall and y the average yield of corn in a certain section; x may be an index of commodity prices and y an index of employment over the same period; we may be interested in a group of school children in which x is their height and y their weight, or x may refer to their reading ability and y to their spelling ability; we may be studying the chance distributions which are obtained in throwing two dice where x is the number obtained in throws of a single die and y is the number obtained in throws of the two dice together.

Example 1. In the following set of selected heights (inches), x = stature of father, y = stature of son.

x	69	70	69	68	70	73	69	67	69	64
y	68	69	72	67	70	71	72	66	71	65

Example 2. (Snedecor.) The following data on twelve trees are adapted from the results of an experiment to test the phenomenon that the injury by codling moth larvae seems to be greatest on apple trees bearing a small crop. Here x = hundreds of fruit on a tree, y = percentage of fruits wormy.

x	15	15	12	26	18	12	8	38	26	19	29	22
y	52	46	38	37	37	34	25	22	22	20	14	

When the given pairs of values are represented by dots locating the points whose rectangular coördinates are (x, y) we obtain a so-called "scatter diagram" (Figure 33). The problem is to determine the degree of association, or correlation as it is called, between the x 's and the corresponding y 's since this indicates the significance of the relationship.

The field of correlation may be thought of as bounded on the one extreme by perfect functional dependence and on the other extreme by complete independence in the probability sense. For example,

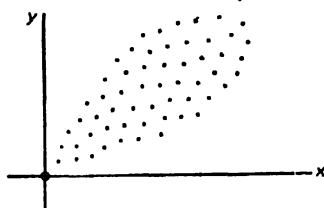


FIG. 33

the pairs of values which satisfy the equation $y = 2x - 5$ do not present a statistical problem. In this case the relationship is defined by a mathematical function $y = f(x)$. Similarly, at the other extreme we would not be concerned with pairs of values which are completely independent in the probability sense, as, for example, the

grades of students in statistics and the heights of their fathers. Two variables are said to be statistically related when they lie between these two extremes of relationship.

The theory of correlation is concerned with a twofold problem: first with measuring the indicated relationship, and secondly with predicting or estimating the average value of y associated with a designated value of x .

2. The Coefficient of Correlation. It is fairly obvious from Figure 33 that with values of x in an assigned interval Δx (Δx small) the corresponding values of y differ considerably. There is said to be positive correlation if, for an assigned x larger than \bar{x} , the mean of the corresponding y values is larger than \bar{y} , and, for values of x smaller than \bar{x} , the mean of the corresponding values of y is less than \bar{y} . On the other hand, as x increases the tendency may be for y to decrease. In this case, for an assigned x larger than \bar{x} the mean of the corresponding y values is less than \bar{y} , and for an assigned x less than \bar{x} the mean of the corresponding y 's is greater than \bar{y} . There is then said to be negative correlation. If, for an assigned x taken at random a corresponding y is no more likely to be above than below \bar{y} , the variables are independent in the statistical or probability sense and there is said to be zero correlation between them.

When the variables are correlated there is a tendency for the dots

Theorem I. *The value of r is independent of the origin of reference and the units of measurement.*

Proof: Let

$$u = \frac{x - x_0}{h}, \quad v = \frac{y - y_0}{k}.$$

Then

$$x = uh + x_0, \quad y = vk + y_0, \quad \sigma_x = h\sigma_u, \quad \sigma_y = k\sigma_v.$$

Substituting in (2) we obtain

$$(4) \quad r = \frac{\frac{1}{N} \sum (u - \bar{u})(v - \bar{v})}{\sigma_u \sigma_v}$$

$$(4a) \quad = \frac{\frac{1}{N} \sum uv - \bar{u}\bar{v}}{\sigma_u \sigma_v},$$

$$\text{where} \quad \sigma_u = \left[\frac{1}{N} \sum u^2 - \bar{u}^2 \right]^{1/2}, \quad \sigma_v = \left[\frac{1}{N} \sum v^2 - \bar{v}^2 \right]^{1/2}.$$

Since (4) and (4a) are independent of the constants x_0 , y_0 , h , and k , the theorem is proved.

This property is of fundamental importance. It means that the units of measurement for the two sets of observed quantities can be chosen independently of each other. If the two sets of quantities are of the same kind, the units need not be the same in both cases; and, what is more important, if the quantities are of different kinds, so that the units are not comparable at all, the coefficient r nevertheless may have a definite meaning. (Of course the value of the coefficient will be affected by a change in the method of measurement of one of the quantities, such as the substitution of an area for a length in estimating the size of an object, or the assignment of different relative weights to the questions on an examination.)

The pairs (x_i, y_i) may be all distinct or there may be repetitions among them. But it is necessary to impose the condition that neither x_i nor y_i shall be constant throughout. This condition is imposed to insure that the denominator shall not vanish in the various formulas for r . *The Algebra of Correlation* — Dunham Jackson. *American Mathematical Monthly*, vol. 31 (1924), pp. 110-121.

When the given values of x and y are large and a computing machine is not available, the computations may be lightened by an appropriate choice of these constants. If only the origin of reference is changed,

then $h = k = 1$, and $u = x - x_0, v = y - y_0$. If the means are taken as the origin of reference by letting $x' = x - \bar{x}$ and $y' = y - \bar{y}$, then $\bar{x}' = \bar{y}' = 0$ and the formula becomes,

$$(5) \quad r = \frac{\frac{1}{N} \sum x'y'}{\left[\frac{1}{N} \sum x'^2 \right]^{1/2} \left[\frac{1}{N} \sum y'^2 \right]^{1/2}}.$$

A subscript notation should be attached to r when there are several series of variates. Thus, r_{xy} for the (x, y) series, r_{xz} for the (x, z) series, r_{12} for the series denoted by (x_1, x_2) , etc.

Example 3. To illustrate the formulas we will compute the value of r for the following data. Here x = Brokers' Loans in billions of dollars and y = *The Annalist's* index of the prices of fifty rail and industrial stocks in 1929. We choose $u = x - 5.00$ and $v = y - 250$.

Month	x	y	u	v	uv	u^2	v^2
J	5.33	248	.33	-2	-0.66	.1089	4
F	5.67	248	.67	-2	-1.34	.4489	4
M	5.65	243	.65	-7	-4.55	.4225	49
A	5.56	249	.56	-1	-.56	.3136	1
M	5.53	235	.53	-15	-7.95	.2809	225
J	5.28	265	.28	15	4.20	.0784	225
J	5.77	282	.77	32	24.64	.5929	1024
A	6.02	303	1.02	53	54.06	1.0404	2809
S	6.35	290	1.35	40	54.00	1.8225	1600
O	6.80	230	1.80	-20	-36.00	3.2400	400
N	4.88	201	-.12	-49	5.88	.0144	2401
D	3.45	206	-1.55	-44	68.20	2.4025	1936
Sums			6.29	0	159.92	10.7659	10678
$\frac{1}{N}$ Sums			.5242	0	13.3267	.8972	889.8333

Computations: $\sigma_u = [.8972 - (.5242)^2]^{1/2} = .79$

$\sigma_v = [889.8333]^{1/2} = 29.83.$

From (4a) we have,

$$r = \frac{13.3267}{(29.83)(.79)} = .57.$$

Experienced computers use calculating machines to great advantage in large-scale computational studies. The following reference is recommended to students who expect to engage in such work: "The Calculation of Correlation Coefficients from Ungrouped Data"—P. S. Dwyer, *Journal of the American Statistical Association*, vol. 35 (1940), pp. 671-673.

Exercises

1. When x' and y' represent deviations from the means,

(a) Show from (1) that $\sum x'y' = Nr\sigma_x\sigma_y$.

(b) Show that $N\sigma_x^2 = \sum x'^2$.

2. Derive formula (3) from (2).

3. Show that (3) may be written as

$$r = \frac{N\sum xy - \sum x \sum y}{[\{N\sum x^2 - (\sum x)^2\} \{N\sum y^2 - (\sum y)^2\}]^{1/2}}.$$

4. Find r for the data of Example 1.

5. Find r for the data of Example 2.

6. The following data represent the ages of husband (x) and wife (y) of twenty couples. Find r using (5). *Ans.* 0.856.

x	22	24	26	26	27	27	28	28	29	30	30	30	31	32	33	34	35	35	36	37
y	18	20	20	24	22	24	27	24	21	25	29	32	27	27	30	27	30	31	30	32

7. In studying a set of pairs of related variates, a statistician has completed the preliminary arithmetic and obtained the following results:

$N = 100$; $\sum x^2 = 1,585,000$; $\sum x = 12,500$; $\sum xy = 1,007,425$; $\sum y^2 = 648,100$; $\sum y = 8,000$. Find \bar{x} , \bar{y} , σ_x , σ_y , r .

8. The table in Exercise 2, page 97, contains the grades made on two tests by twenty-five students in mathematics. Find r for these data. *Ans.* 0.786.

9. Suggest examples of negative correlation.

10. In the following anthropometric measurements on a random sample of twenty male freshmen, taken from the Physical Education Department,

x	y	z	x	y	z
68.5	33.6	148	65.3	33.0	136
67.2	35.0	144	65.1	34.0	144
67.7	30.2	145	64.8	37.3	170
63.8	30.0	108	69.6	33.4	154
69.9	33.0	130	68.2	31.5	122
64.7	31.0	112	68.8	32.0	141
68.4	33.0	134	72.3	35.0	159
66.4	30.2	112	67.8	33.7	134
69.1	33.3	143	71.3	31.5	136
71.0	32.3	136	63.5	33.6	126

x represents height, y represents chest measurement, both measurements being taken to the nearest tenth of an inch, and z represents weight to the nearest pound. Find the coefficient of correlation (a) between x and y , (b) between x and z , (c) between y and z .

4. Regression. The properties of r can be studied by fitting a line to the scatter diagram in such a way as to make the sum of the squares of the vertical distances from the points to the line a minimum.

When such a line is referred to the point (\bar{x}, \bar{y}) as origin, we have seen (§9, Chapter VII) that its equation is $y' = m_1 x'$ where

$$m_1 = \frac{\sum x'y'}{\sum x'^2}$$

and $x' = x - \bar{x}$, $y' = y - \bar{y}$. This value of m_1 may easily be expressed in terms of r and the standard deviations, as follows:

$$m_1 = \frac{Nr\sigma_y\sigma_x}{N\sigma_x^2} = r \frac{\sigma_y}{\sigma_x}.$$

Therefore, the equation of our line, referred to a system of axes whose origin is at the means of the variates, is

$$(6) \quad y' = \frac{\sigma_y}{\sigma_x} r x'.$$

This is called the *regression line of y on x* . The term *regression* was used first by Galton in studying inheritance of stature. He found that offspring of abnormally tall or short parents tend to "step back" or "regress" to the ordinary population height. However, as now used, regression line has no reference to biometry, but is merely a convenient term.

By fitting a line $x' = m_2 y'$ to the points of the scatter diagram in such a way that the sum of the squares of the horizontal distances from the points to the line shall be a minimum, it is possible to deduce a second regression line (the regression line of x on y) whose equation, referred to (\bar{x}, \bar{y}) , is

$$(7) \quad x' = \frac{\sigma_x}{\sigma_y} r y'.$$

Note that (7) cannot be obtained by solving for x' in (6). The two regression lines will coincide if, and only if, $r = \pm 1$. From

the equations of the regression lines it is evident that if $r > 0$, an increase in the one variable tends to accompany an increase in the other; if $r < 0$, an increase in the one will be accompanied by a decrease in the other.

Equations (6) and (7) are usually expressed in terms of the original variables x and y instead of the deviations x' and y' . It is obvious that they may be written as

$$(8) \quad y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

and

$$(9) \quad x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

when referred to the origin of x and y .

Equation (8) may be used to estimate values of y corresponding to designated values of x . Similarly, from equation (9) we may estimate x for designated values of y . It would be appropriate to use (8) as a predicting equation when the variation in y is caused or controlled by the variation in x ; (9) would be used when the variation in x is caused or controlled by the variation in y .

The quantity $m_1 = r(\sigma_y/\sigma_x)$ is called the regression coefficient of y on x , being the variation in y corresponding to a unit change in x . Likewise, $m_2 = r(\sigma_x/\sigma_y)$ is called the regression coefficient of x on y . Thus the numerical value of r is given by $(m_1 m_2)^{1/2}$ but its sign must be that which is common to the two regression coefficients. The following quotation from Snedecor (reference 13, list p. 6) sheds light on the distinction between regression and correlation.

The point of interest here is that r is the geometric mean of the two regression coefficients. In ordinary units of measurement, therefore, r is an average of the two regression coefficients used in (i) estimating y from x and (ii) estimating x from y . This serves to clarify the relation of the two coefficients, correlation and regression, in measuring relationship. The latter is the appropriate one if one variable, y , may be designated as dependent on the other, x . Values of y may be partly controlled or caused by x , as when the available amounts of some glandular secretion cause differences in the sizes of organisms. Or, y may be subsequent to x , as weight gain in nutrition experiments follows the measurement of initial weight. In such cases, the regression of y on x is usually the statistic that furnishes the information desired. It is then appropriate to attempt to estimate the value of y from a knowledge of the corresponding value of x . Correlation, on the other hand, is the appropriate measure of the relation between

two variates like statures of husband and wife. The two heights are known to be associated through some complex of social and biological causes, but neither may be looked upon as a consequence of the other. In this sense correlation is a two-way average of relationship, while regression is directional. Of course, there are many variables whose relationship may be studied by means of either correlation or regression, or both. It is necessary only to keep clearly in mind the character of the relation being considered.

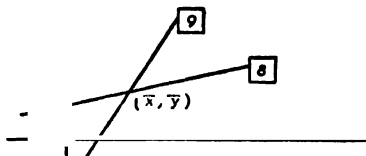


FIG. 35

Geometrically, m_1 is the slope of line (8) and $1/m_2$ is the slope of line (9). The two lines intersect at (\bar{x}, \bar{y}) .

Exercises

- Derive the equation of the line of regression of x on y as suggested above.
- Find the equations of both lines of regression for Exercise 6 (page 176), and plot them. *Ans.* $y = .888x - .64$
 $x = .825y + 8.55$.
- Using the appropriate equation, find the estimated values of y corresponding to the given values of x , for Exercise 6 (page 176).
- Given the following results for the heights and weights of 1000 men students:
 $\bar{y} = 68.00$ in., $\bar{x} = 150.00$ lbs., $r = .60$,
 $\sigma_y = 2.50$ in., $\sigma_x = 20.00$ lbs.

John Doe weighs 200 lbs., Richard Roe is five feet tall.

Estimate the height of Doe from his weight, and the weight of Roe from his height.

Ans. Doe's height = 71.75 in.

Roe's weight = 111.6 lbs.

- (a) Given the following:

$$\sum x = 150,000, \sum x^2 = 22,725,000, \sum xy = 10,522,500,$$

$$\sum y = 70,000, \sum y^2 = 4,936,000, N = 1000.$$

Find \bar{x} , \bar{y} , σ_x , σ_y , r , and the lines of regression.

- (b) Suppose the data in (a) refer to the weight in pounds (x) and the height in inches (y) of a sample of 1000 policemen. Suppose Paul Private weighs 160 pounds and Saul Sergeant is 6 feet tall. Estimate the height of Private and the weight of Sergeant.

5. The Standard Error of Estimate. The average concentration of the points around the regression line of y on x may be measured by the expression $\frac{1}{N} \sum d^2$ where d is the difference between an observed y and the y obtained from the regression line. The value of

$\frac{1}{N} \sum d^2$ will be denoted by S_y^2 , and S_y is called the standard deviation of the errors of estimate, or more briefly the *standard error of estimate*. The errors of estimate are the deviations of the observed values of y from the corresponding estimated y 's. (Or to describe them another way, they are the deviations of the sample y 's from the assumed population y 's. It can be shown that $S_y^2 = \sigma_y^2(1 - r^2)$. To prove this we may write the sum of the squares of the deviations in the form:

$$\begin{aligned} NS_y^2 &= \sum \left(y' - r \frac{\sigma_y}{\sigma_x} x' \right)^2 = \sum y'^2 - 2r \frac{\sigma_y}{\sigma_x} \sum x'y' + r^2 \frac{\sigma_y^2}{\sigma_x^2} \sum x'^2 \\ &= N\sigma_y^2 - 2Nr^2\sigma_y^2 + Nr^2\sigma_y^2 = N\sigma_y^2(1 - r^2). \end{aligned}$$

Hence, we have

$$(10) \quad S_y^2 = \sigma_y^2(1 - r^2)$$

and

$$(10a) \quad S_y = \sigma_y(1 - r^2)^{1/2}.$$

An analogous consideration of the differences between the x 's and the regression line of x on y gives for the square of the standard error of estimate of the x 's

$$(11) \quad S_x^2 = \sigma_x^2(1 - r^2).$$

6. Properties of the Correlation Coefficient and Standard Error of Estimate. Certain properties of r may now be deduced. It is obvious from (10) that $|r| \leq 1$ because both the left member and σ_y^2 are positive or zero. Therefore,

$$-1 \leq r \leq 1.$$

If the points all lie exactly on the regression line, the left member of (10) vanishes and $r = \pm 1$. There is then said to be perfect linear correlation, since the relation between x and y is given exactly by a linear function. A large numerical value of r means that the regression lines are close to coincidence and the points in a scatter diagram cluster closely around the regression lines.

When the regression lines (8) and (9) are expressed in standard units, they become respectively

$$(12) \quad t_y = rt_x$$

and

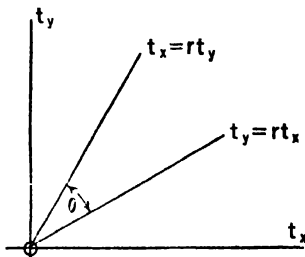
$$(13) \quad t_x = r t_y \quad \text{or} \quad t_y = \frac{1}{r} t_x$$

where

$$t_x = \frac{x - \bar{x}}{\sigma_x} \quad \text{and} \quad t_y = \frac{y - \bar{y}}{\sigma_y}.$$

In this form we see at once that as one variable t_x increases, the other variable t_y increases (or decreases) to an extent that depends upon r . Thus r measures co-variation in the variables when they are expressed in comparable units and when regression is linear.

In standard units, r is the slope of line (12) and $1/r$ is the slope of line (13). When $r = 0$, the regression equations become $t_y = 0$ and $t_x = 0$ in standard units or $y = \bar{y}$ and $x = \bar{x}$ in the original units. These are also the equations of the coördinate axes. Therefore, when $r = 0$ the regression lines are perpendicular to each other and coincide with the t_x and t_y axes. When $r = 1$ the regression equations become identical and the two lines coincide in quadrants I and III. Similarly, when $r = -1$ they coincide in quadrants II and IV. In each case the coincident lines bisect the quadrants if the equations are expressed in standard units, but not otherwise unless $\sigma_y = \sigma_x$. The angle θ between the regression lines varies from 0° to 90° as r varies from one to zero.



When there is no correlation between x and y then $r = 0$, and the variables are said to be independent in the statistical sense. On the other hand, when $r = 0$, it is not necessarily true that the variables are statistically independent. Indeed there may be a high correlation¹ with non-linear regression when $r = 0$. (Non-linear regression will be considered in §21.) Incidentally, the phrase "independent variables" in the statistical sense should not be confused with the phrase "independent variables" which is used in the ordinary sense of analysis to designate the variables on which a specified function depends. However, the two usages, though quite distinct, are not fundamentally contradictory, since functional dependence can be regarded as a limiting case of statistical dependence.

¹ See H. L. Rietz, *On Functional Relations for which the Coefficient of Correlation is Zero*. *Journal American Statistical Association*, vol. 16, 1919, pp. 472-476.

For an appreciation of the use of S_y in passing judgment upon the precision to be expected in estimating values of y by means of the regression equation of y on x , it is instructive to consider simultaneously the meanings of (8) and (10a) as $|r|$ varies from 0 to 1. When $r = 0$, (8) becomes $y = \bar{y}$ which means that the best estimate of y for *any* value of x is the mean of the y -distribution. In other words, knowledge of x is of no value in predicting y . When $r = 0$ in (10a), $S_y = \sigma_y$. This is to be expected since the dispersion S_y about the line $y = \bar{y}$ is the same as the dispersion σ_y of the given y 's about their mean. But as $|r|$ increases from 0 to 1, S_y decreases from σ_y to 0. Graphically, the meaning of this improvement in S_y in comparison

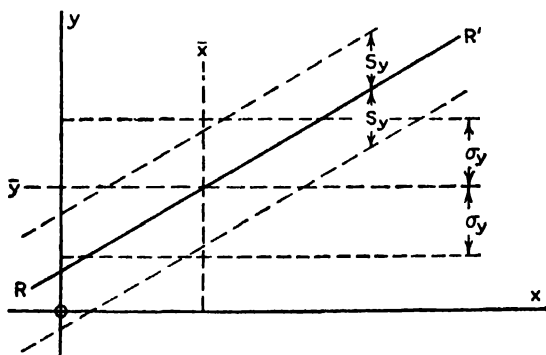


FIG. 36 — FOR A FIXED VALUE OF σ_y , S_y DECREASES IN PROPORTION TO $(1 - r^2)^{1/2}$ AS r INCREASES

with σ_y , as r increases, is shown in Figure 36 where parallel lines are drawn at a vertical distance of S_y on either side of the regression line RR' . For a given value of $|r| \neq 0$ this strip encloses the average dispersion about the line. The strip on either side of $y = \bar{y}$ at a distance of σ_y from it encloses the average dispersion about the line when $r = 0$. As $|r|$ increases from 0, the line rotates from the horizontal position of $y = \bar{y}$ to the terminal position it would have when $|r| = 1$, and at the same time S_y decreases toward 0. Formula (10a) tells us that as $|r|$ thus increases, S_y decreases from σ_y in proportion to $(1 - r^2)^{1/2}$.

A similar analysis could be made concerning the line of regression (9) of x on y which rotates from the vertical position $x = \bar{x}$ when $|r| = 0$ to meet and coincide with line (8) when $|r| = 1$. As line (9) rotates, S_x decreases from σ_x to 0 in proportion to $(1 - r^2)^{1/2}$ as r increases.

As $|r| \rightarrow 1$, (12) and (13) rotate toward each other at equal angular velocities. When they are coincident their slope is ± 1 . Lines (8) and (9) rotate at angular velocities which are proportional to $m_1 = \tan \alpha$ and $m_2 = \tan \beta$, respectively, when m_1 and m_2 are defined in §4. Their slope at coincidence is $\pm \sigma_y/\sigma_x$. For line (12) it can be shown that

$$(14) \quad \frac{1}{N} \sum \delta^2 = 1 - r^2$$

where δ is the difference between an observed value of t_y and the ordinate obtained from (12) for the corresponding value of t_x . Thus,

$$\begin{aligned} \frac{1}{N} \sum \delta^2 &= \frac{1}{N} \sum (t_y - r t_x)^2 \\ &= \frac{1}{N} \sum t_y^2 - \frac{2r}{N} \sum t_x t_y + \frac{r^2}{N} \sum t_x^2 \\ &= 1 - 2r^2 + r^2 \\ &= 1 - r^2. \end{aligned}$$

This result would also be apparent from the derivation of (10) since $\delta = d/\sigma_y$ where d refers to residuals in units other than standard units.

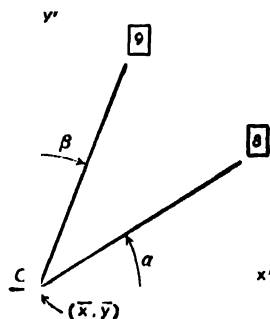
It is obvious from (14) that the maximum value of $\frac{1}{N} \sum \delta^2$ is unity.

Therefore, adopting

$$(15) \quad 1 - \frac{1}{N} \sum \delta^2$$

as a measure of goodness of fit, we see from (14) and (15) that r^2 is a measure of the goodness of fit of (12) to the points of the scatter diagram expressed in standard units. By an analogous argument a similar conclusion concerning (13) can be made.

7. Further Discussion. Given a set of N pairs of x and y correlated values. Suppose the necessary constants are evaluated to obtain the regression equation (8). Then if the given values of x are substituted in this equation, a set of estimated y 's, say \hat{y} , will be



obtained. The mean, ${}_E\bar{y}$, of these estimated y 's is the same as the mean of the observed y 's. The proof is as follows. From (8) we have

$${}_E y = \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x}).$$

Then

$${}_E\bar{y} = \frac{1}{N} \sum_1^N \bar{y} + r \frac{\sigma_y}{\sigma_x} \frac{1}{N} \sum_1^N (x_i - \bar{x}).$$

But $\sum_1^N (x_i - \bar{x}) = 0$ by Theorem VI, Chapter III. So ${}_E\bar{y} = \bar{y}$.

We now state the following theorem.

Theorem II. *The variance, $\sigma_{{}_E y}^2$, of the estimated y 's equals $r^2 \sigma_y^2$.*

Proof: By definition,

$$\sigma_{{}_E y}^2 = \frac{1}{N} \sum ({}_E y_i - {}_E\bar{y})^2.$$

From the above discussion, $({}_E y_i - {}_E\bar{y})$ is the same as $(y_i - \bar{y})$ which is given by (8). So

$$\begin{aligned} \sigma_{{}_E y}^2 &= \frac{1}{N} \sum_1^N \left[r \frac{\sigma_y}{\sigma_x} (x_i - \bar{x}) \right]^2 \\ &= r^2 \frac{\sigma_y^2}{\sigma_x^2} \frac{1}{N} \sum_1^N (x_i - \bar{x})^2. \end{aligned}$$

Hence

$$(16) \quad \sigma_{{}_E y}^2 = r^2 \sigma_y^2.$$

From this theorem and (10) we obtain

$$(17) \quad S_y^2 = \sigma_y^2 - \sigma_{{}_E y}^2.$$

This relation helps to clarify the meaning of r and of S_y . It is conventional to call $\sigma_{{}_E y}^2$ the variance in y which can be *explained* from knowledge of x ; that is, which the regression of y on x accounts for. (In the language of some writers, $\sigma_{{}_E y}^2$ measures the variation of regression about the mean.) Therefore, (17) shows that S_y^2 is the variation in y after the accompanying variation in x is duly discounted. S_y^2 is sometimes called the *residual* variance because it measures the variation in the dependent variable y which knowledge of x fails to account for. This relation can be depicted geometrically by the sides of a right triangle. To standardize the representation we can take $\sigma_y = 1$ as the diameter of a semicircle within which

is inscribed the right triangle, as in Figure 37. In the figure, $\cos \theta = \sigma_{Ey}/\sigma_y$. So from (16) we have $\cos \theta = r$. The particular values of θ in the figure, found from a table of cosines, are $\theta = 36^\circ 52'$ when $r = .8$, and $\theta = 25^\circ 50'$ when $r = .9$. When $r = 1$, then $\sigma_{Ey} = \sigma_y$ and the regression of y on x accounts for all the variation in y .

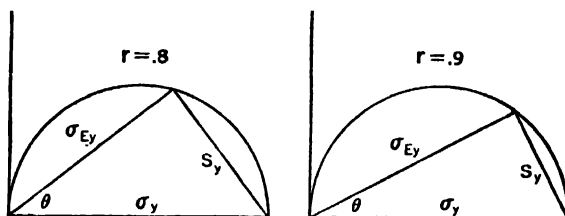


FIG. 37

Theorem III. *The correlation between observed and estimated values of y is the same as that between the observed values of x and y .*

Proof: We are to show that

$$\frac{\frac{1}{N} \sum y_E y - \bar{y}_E \bar{y}}{\sigma_{Ey} \sigma_y}$$

reduces to one of the formulas for r . Substituting the values for Ey , $E\bar{y}$, σ_{Ey} into the above expression and simplifying, we obtain (3). The details of the proof are left to the student as an exercise.

8. Coefficient of Alienation. A measure of the failure to improve estimates of y from knowledge of correlation is given by

$$(18) \quad k = (1 - r^2)^{1/2}.$$

It is sometimes called the coefficient of alienation. Incidentally, it is interesting to observe that the functional relation between k and r is shown, graphically, by a semicircle of unit radius, *i.e.*,

$$f(r) = (1 - r^2)^{1/2}.$$

The formula

$$k' = 1 - (1 - r^2)^{1/2}$$

may be called the *improvement factor* because it shows the decrease in S_y/σ_y as $|r|$ increases. It is clear that

$$k^2 = \frac{1}{N} \sum \delta^2 = \frac{\sigma_y^2}{\sigma_y^2} \quad \text{and} \quad k' = 1 - k.$$

Table 31 gives¹ values of k and k' for values of r . With no knowledge of correlation, the best estimate of an individual y is \bar{y} . Values of k' for assigned r 's show how much better than this guess is the estimate of an individual y value with knowledge of correlation. For example, when $r = .5$ the column headed k in Table 31 shows that the standard error S_y is about 87% of σ_y . Or, from the k' column, S_y has been reduced only 13% from what it would have been if \bar{y} had been used for prediction purposes. The third column thus shows how the prediction value of r varies with r . Thus as $|r|$ decreases from 1 to .8, S_y/σ_y increases from 0 to 60%. Or from another point of view, as $|r|$ increases from 0 to .8, the error of estimate is improved by only 40%. A correlation of $r = .9$ permits prediction of individual y 's only 56% better than a mere guess based on the mean.

It is fairly obvious that we cannot, with any considerable degree of reliability, predict from ordinary values of r an individual y for an assigned x . However, with a large N , we can give a very reliable prediction of the mean of y values that correspond to an assigned value of x . This can best be explained from a correlation table which is used when N is large and which will be explained in the next section.

TABLE 31 — VALUES OF r AND THE CORRESPONDING VALUES OF k AND k'

r	k	k'
.1	.995	.005
.2	.980	.020
.3	.954	.046
.4	.917	.083
.5	.866	.134
.6	.800	.200
.7	.714	.286
.8	.600	.400
.9	.436	.564
.92	.392	.608
.94	.341	.659
.96	.280	.720
.98	.198	.811
1.00	0.000	1.000

¹ Constructed from a table of sines and cosines. Letting $r = \cos \theta$, $\sin \theta = (1 - \cos^2 \theta)^{1/2} = (1 - r^2)^{1/2}$.

Exercises

1. Given the following correlated data:

x	8	6	4	7	5
y	9	8	5	6	2

- Compute the correlation coefficient.
 - Find the regression line of y on x .
 - Find the estimated values of y corresponding to the given values of x .
 - Compute the standard error S_y of predictions in two different ways.
- Ans.*

$$r = \frac{2.4}{\sqrt{2}\sqrt{6}} = .69, \quad m_1 = r \frac{\sqrt{6}}{\sqrt{2}} = 1.2, \quad S_y = \sqrt{3.12} = 1.76.$$

Note. In practical work, it is never worth while calculating a correlation coefficient for so few observations. These fictitious data are given solely as an exercise on which the student can test his knowledge of the methodology.

- Prove that the ratio of variance of the estimated y 's (taken about their mean) to the variance σ_y^2 of the given y 's is equal to r^2 .
- If $S_y^2/\sigma_y^2 = 1 - r^2$ is the percentage of the total variance of y uncontrolled by knowledge of x , what is the remaining percentage, determined by or calculable from knowledge of x ?
- What equation is the equivalent mathematical statement for the following words?

If the respective deviations in each series, x and y , from their means were expressed in units of standard deviations—that is, if each were divided by the standard deviation of the series to which it belongs—and plotted to a scale of standard deviations, the slope of a straight line best describing the plotted points would be the correlation coefficient r .

- Given the standard deviations σ_x and σ_y of two distributions of correlated variates:
 - What is the standard error in estimating y from x if $r = 0$?
 - By how much is S_y in (a) reduced if r is increased to .25?
 - How large must r be in order that S_y be one-half as large as in (a)?
 - What must r be in order that S_y be reduced to one-third its value in (a)?
 - At what value of r is S_y reduced to zero?
 - For any value of r , what is the ratio between the standard error of estimating y from x and the standard deviation of the y -distribution?
- Evaluate the following statements:
 - A correlation coefficient less than zero indicates an absence of linear relationship.
 - A correlation coefficient of $r = .6$ indicates twice as close relationship as a coefficient of $r = .3$.

7. If all the points lie exactly on the regression line of y on x , show that $S_y^2 = 0$ and hence that $r = \pm 1$.
8. Show that S_y^2 may be computed by means of the relation

$$NS_y^2 = \sum y'^2 - \frac{(\sum x'y')^2}{\sum x'^2}$$

where the primes denote deviations from the means.

9. (For analytics students.) Show that the tangent of the angle from line (8) to line (9) is

$$\tan \theta = \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \left\{ \frac{1 - r^2}{r} \right\}$$

and from line (12) to line (13) is

$$\tan \theta = \frac{1 - r^2}{2r}$$

What is the value of θ when $r = 1$; when $r = 0$?

10. The least-squares criterion of best fit requires that $\sum \delta^2$ be a minimum, where δ is the distance between the line and a point. Three cases arise depending on whether

Case I, δ is measured parallel to the y -axis,

Case II, δ is measured parallel to the x -axis,

Case III, δ is measured perpendicular to the line.

We have seen that Case I yields line (12) and that Case II yields line (13).

In Case III the line has no universally accepted name but it may be called the "geometrically best-fitting line."

(For calculus students.) For Case III prove the following:

(a) In standard units, the equation of the line is

$$t_y = t_x \quad \text{if} \quad r > 0$$

$$\text{and} \quad t_y = -t_x \quad \text{if} \quad r < 0.$$

Solution. Let the equation of the required line be

$$t_y = mt_x + k.$$

Then by analytics,

$$\begin{aligned} \frac{1}{N} \sum \delta^2 &= \frac{1}{N} \sum \left(\frac{mt_x + k - t_y}{\sqrt{1 + m^2}} \right)^2 \\ &= \frac{m^2 + k^2 + 1 - 2mr}{1 + m^2}. \end{aligned}$$

To make this a minimum, first put $k^2 = 0$. Call the result $f(m)$. Then

$$f(m) = \frac{m^2 + 1 - 2mr}{1 + m^2},$$

$$f'(m) = \frac{2m^2r - 2r}{(1 + m^2)^2},$$

$$f''(m) = \frac{4mr(3 - m^2)}{(1 + m^2)^3}.$$

The second derivative will be positive when m and r have the same sign.

Since $f(m)$ is a minimum when $m = \pm 1$, we are to take $m = 1$ when $r > 0$ and $m = -1$ when $r < 0$.

(b) If $r = 0$, all lines (for which $k^2 = 0$) fit equally well. *Hint.* If $r = 0$, $f(m) = 1$.

(c) $\frac{1}{N} \sum s^2 = 1 - |r|$. *Hint.* What is the value of $f(m)$ when $m = \pm 1$?

Note that $|r| = +r$, if $r > 0$ and $|r| = -r$, if $r < 0$.

(d) Goodness of fit is measured by $|r|$.

(e) When $r = .6$ the fit is twice as good as when $r = .3$.

11. The following query and answer appeared in *Biometrics Bulletin*, vol. 1, no. 3, pp. 36-37. "Research" assignment: Investigate the references cited in the answer and justify the procedure which is recommended (under the given hypothesis).

Query. A problem that has bothered me is the fitting of regression lines when their position is restricted in some way. For example, suppose a test is made of the relationship between the number of fish caught in a body of water and the average number which can be caught out of it, with a standard amount of fishing. In fitting a regression line to such data, we know that the point (0, 0) must fall on the line, since if no fish are present certainly none will be caught. In other words, we have one point which is free from sampling error. The unique importance of this point will, it seems to me, make observations in its neighborhood of relatively less importance than observations at a distance from it, where there is no fixed guide-post. Do you know of any treatment of situations of this sort, by which the best straight (or curved) line could be fitted to data where there is one point which *must* be satisfied? The standard deviation from regression ("standard error of estimate") and the standard error of the regression would also be available. Or are these concepts pertinent in such a question?

Answer. Deming (§15 and §11 of reference 4) gives both a general method and some particular solutions of your problem. Snedecor (reference 6) opens his Chapter 6 with an illustration of the simple case in which x is measured without error and the variance of y is constant for all values of x .

Observations in the neighborhood of (0, 0) may or may not be of less importance than those at greater distances; it depends on the variance of y . One often finds that this variance increases with x . In fact, there are many situations in which it seems reasonable to suppose that in the sampled population the standard deviation of y is directly proportional to x . If you think this hypothesis is suitable in your fishing, the appropriate method is to calculate the ratios x/y where x is the number of fish caught and y is the total number of fish, then apply to them the statistical procedure suitable for a single variate. — George W. Snedecor.

9. Correlation Table. When the sample to be studied is large, it is more convenient to replace the scatter diagram by a correlation

table. We may divide the xy -plane into rectangles of convenient size, and all points of the scatter diagram falling within any rectangle are thought of as being concentrated at the center of this rectangle. A number is then written within the rectangle to designate the number of points at its center. A correlation table is therefore a two-way frequency table exhibiting the frequencies in each class interval.

TABLE 32

		65 - 69	70 - 74	75 - 79	80 - 84	85 - 89	90 - 94	95 - 99	
	$y \backslash x$	67	72	77	82	87	92	97	$f(y)$
90-94	92				1	2	3	1	7
85-89	87			1	3	8	1	5	18
80-84	82	4	4	6	4	9	1		28
75-79	77	3	3	7	6	4			23
70-74	72	2	3	5	6	1	1		18
65-69	67	3	2						5
60-64	62	1							1
	$f(x)$	13	12	19	20	24	6	6	100

Suppose Table 32 is constructed in this way for a set of average daily grades (x) and final examination grades (y) of 100 students. When the data have been thus grouped into classes, the class marks are regarded as the variate values. Thus in Table 32 there are 9 students whose daily grades are 87 and whose final examination grades are 82. The last column labeled $f(y)$ represents the distribution of y variates and the last row labeled $f(x)$ represents the distribution of x variates. A correlation table is thus a bivariate distribution. In Table 32 the width of the class interval is the same for x and y , but of course this is not generally the case.

10. Notation. In order to compute r from a correlation table it will be necessary to develop new notation. Since we are now dealing with frequencies in both the x -direction and the y -direction, we will distinguish between them by $f(x)$ and $f(y)$. To be sure, this has the disadvantage of being the same symbol as that for function, but from the context no ambiguity should arise.

Generalizing, a correlation table is of the following form:

$y \backslash x$	x_1	x_2	x_3	—	$\downarrow x$ —	—	x_n	$f(y)$
y_n								
y_{n-1}			$f(x, y)$					$\sum_x f(x, y)$
\vdots								
$\bar{y} \rightarrow \bar{1}$				(\bar{x}, \bar{y})				
\vdots								
y_1								
$f(x)$		$\sum_y f(x, y)$					N	$\sum_y \sum_x f(x, y)$
		y					$\downarrow y$	$\sum_y \sum_x f(x, y)$

The rectangles containing the frequencies are called cells. The frequency in a typical cell is denoted by $f(x, y)$, meaning the frequency in the cell whose coördinates are x and y , where x and y are the mid-values of the class intervals. Both columns and rows are sub-distributions of the total frequency N . Each column is a frequency distribution of y 's corresponding to a mid- x value. Similarly, each row is a frequency distribution corresponding to a mid- y value. The sum along any row is denoted by $\sum_x f(x, y)$, being the sum of the frequencies in the (x, y) cells in the x -direction. Since the marginal total for any row is the total frequency corresponding to a given value of y , it is therefore written in the column headed $f(y)$. Thus, in Table 32, for $y = 92$,

$$\sum_x f(x, y) = \sum_x f(x, 92) = 1 + 2 + 3 + 1 = 7.$$

Similarly, $\sum_y f(x, y)$ denotes a summation in the y -direction of all the entries in a column, corresponding to a fixed value of x , so it denotes an entry in the bottom row which contains the $f(x)$ frequencies. Thus, for $x = 67$

$$\sum_y f(67, y) = 4 + 3 + 2 + 3 + 1 = 13.$$

Summarizing,

$$(19) \quad \sum_x f(x, y) = f(y); \quad \sum_y f(x, y) = f(x). \quad .$$

With regard to N , we may obtain it from a correlation table in three ways: (1) by adding the entries across the rows and then totaling the resulting sums in the marginal column labeled $f(y)$; (2) by adding the entries along the columns and then totaling the results in the marginal row labeled $f(x)$; (3) by adding the entries in the cells in any order whatsoever. Hence, the following notation,

$$(20) \quad \sum_y \sum_x f(x, y) = \sum_x \sum_y f(x, y) = \sum_{x, y} f(x, y) = N,$$

will denote, respectively, the above-named procedures or orders in summing. From (19) and (20) we have

$$(21) \quad N = \sum_y f(y) = \sum_x f(x) = \sum_{x, y} f(x, y).$$

We may call $f(x)$ and $f(y)$ the marginal distributions of x and y , respectively. A correlation table with cell frequencies $f(x, y)$ uniquely determines the marginal totals $f(x)$ and $f(y)$. The converse, however, is false. For example, we might replace the four cell frequencies in the upper right-hand corner of Table 32 by the cell

frequencies

2	2
2	4

 without disturbing the marginal totals.

11. Means and Variances. We will now express the means in terms of this notation, taking first the mean of x 's. From the fundamental definition, we must multiply each x by its corresponding frequency in the cells and sum the results, taking the products in any order whatsoever. Hence,

$$\bar{x} = \frac{1}{N} \sum_{x, y} x f(x, y).$$

This may also be written

$$\bar{x} = \frac{1}{N} \sum_x \sum_y x f(x, y) = \frac{1}{N} \sum_x x \sum_y f(x, y) = \frac{1}{N} \sum_x x f(x).$$

Observe that the x may be moved to the left of \sum_y in the second expression because x is treated as a constant in a summation performed with respect to y .

Similarly, we have,

$$\begin{aligned}\bar{y} &= \frac{1}{N} \sum_{x,y} yf(x,y) = \frac{1}{N} \sum_y \sum_x yf(x,y) \\ &= \frac{1}{N} \sum_y y \sum_x f(x,y) = \frac{1}{N} \sum_y yf(y).\end{aligned}$$

The student will observe that the last expression for the mean in each case is identical with that given for a frequency distribution of one variable, when allowance is made for the necessity of distinguishing between variables.

Any column is an x array of y 's, so the symbol \bar{y}_x is appropriate for the mean of a column. Similarly, \bar{x}_y denotes the mean of a y array of x 's, i.e., of a row. We may now state the following theorem.¹

Theorem IV. *The mean \bar{y} for the whole table (in the y -direction) is equal to the mean of the values \bar{y}_x for the several columns when each \bar{y}_x is weighted with the frequency in that column.*

Proof: We are required to show that

$$\frac{1}{N} \sum_x f(x) \bar{y}_x = \bar{y}$$

where

$$\bar{y}_x = \frac{1}{f(x)} \sum_y yf(x,y).$$

Upon substituting in the first equation the value of \bar{y}_x as given by the second equation, we have

$$\frac{1}{N} \sum_x \sum_y yf(x,y) = \frac{1}{N} \sum_{x,y} yf(x,y) = \bar{y}.$$

It is suggested that the student state and prove a similar theorem concerning \bar{x} .

In this new notation, the definitions of the variances becomes

$$\begin{aligned}\sigma_x^2 &= \frac{1}{N} \sum_{x,y} (x - \bar{x})^2 f(x,y) \\ &= \frac{1}{N} \sum_x x^2 f(x) - \bar{x}^2;\end{aligned}$$

¹ This is actually the same as Theorem IX on page 45, but it seems worthwhile to state and prove it in the new notation.

$$\begin{aligned}\sigma_y^2 &= \frac{1}{N} \sum_{x,y} (y - \bar{y})^2 f(x, y) \\ &= \frac{1}{N} \sum_y y^2 f(y) - \bar{y}^2.\end{aligned}$$

Exercises

1. Evaluate the following expressions in Table 32.

(a) For $x = 82$,

$$\sum_y f(x, y), \quad \sum_y y f(x, y), \quad f(x), \quad \bar{y}_x.$$

(b) For $y = 87$,

$$\sum_x f(x, y), \quad \sum_x x f(x, y), \quad f(y), \quad \bar{x}_y.$$

2. Refer to Table 27 (Chapter V) and let x be the number of a column. Express the answers in the third and second lines from the bottom of the table in terms of the notation of this section. Thus for $x = 1$,

$$\bar{y}_x = \frac{1}{f(x)} \sum_y y f(x, y) = \frac{1}{7} [85 + (75)2 + (65)2 + (55)2] = 67.86.$$

12. Computation of Means. Just as in the case of a one-way frequency distribution it was found convenient to choose an arbitrary origin and take the class interval as the unit, so we now do likewise. Let

$$(22) \quad u = \frac{1}{h} (x - x_0); \quad \text{i.e.,} \quad x = uh + x_0.$$

Hence,

$$(23) \quad \bar{x} = \bar{u}h + x_0$$

where

$$\bar{u} = \frac{1}{N} \sum_u u f(u).$$

Likewise, let

$$(24) \quad v = \frac{1}{k} (y - y_0); \quad \text{i.e.,} \quad y = vk + y_0.$$

whence

$$(25) \quad \bar{y} = \bar{v}k + y_0,$$

where

$$\bar{v} = \frac{1}{N} \sum_v v f(v).$$

Then a suitable form for computing the means of the x 's and y 's is as follows:

	u	-3	-2	-1	0	1	2	3	$f(y)$	$v f(v)$
v	$y \backslash x$	67	72	77	82	87	92	97	$f(v)$	
3	92				1	2	3	1	7	21
2	87			1	3	8	1	5	18	36
1	82	4	4	6	4	9	1		28	28
0	77	3	3	7	6	4			23	0
-1	72	2	3	5	6	1	1		18	-18
-2	67	3	2						5	-10
-3	62	1							1	-3
$f(x)=f(u)$		13	12	19	20	24	6	6	100	54
$u f(u)$		-39	-24	-19	0	24	12	18	-28	

Computations:

$$\bar{u} = \frac{1}{N} \sum u f(u) = \frac{-28}{100} = -.28,$$

whence

$$\bar{x} = 82 + 5(-.28) = 80.6.$$

$$\bar{v} = \frac{1}{N} \sum v f(v) = .54,$$

whence

$$\bar{y} = 77 + 5(.54) = 79.7.$$

In the table $f(v) = f(y)$ and $f(u) = f(x)$ because u and v are merely different ways of describing the cells but in no way change the frequencies in those cells.

13. Computation of r . In the expressions of §10 and §11 the (u, v) coördinates could have been used instead of (x, y) . The use of the former simplifies the computation of r . A preliminary discussion of certain expressions will help in understanding the formula for r to be used for a correlation table. Let us consider first the following expression:

$$(a) \quad \sum u v f(u, v).$$

This means: multiply the f in each cell by the u and v coördinates of that cell and add the results, proceeding from cell to cell over the whole table in any order whatsoever. But it may be more con-

venient to proceed in a definite order, say down the columns. Then (a) becomes

$$(b) \quad \sum_u \sum_v uvf(u, v) = \sum_u u \sum_v vf(u, v).$$

The expression $\sum_v vf(u, v)$ in the right member of (b) means: for any u (i.e., for any column), multiply each f by its own v and add the results. Let us denote this sum by V . Then the right member of (b) means: multiply the V for each column by the u of that column and add the results, proceeding from column to column (i.e., summing in the u -direction). We may also obtain the same result as in (a) by proceeding along the rows. Thus (a) may be written

$$(c) \quad \sum_v \sum_u uvf(u, v) = \sum_v v \sum_u uf(u, v).$$

The expression $\sum_u uf(u, v)$ means: for any v (i.e., for any row), multiply each f in the row by its own u and add the results. If we call this sum U , then the right member of (c) means: multiply the U for each row by the v for that row and add the results, proceeding from row to row (i.e., summing in the v -direction).

We are now ready to derive the formula for r .

Since we are now dealing with a frequency distribution, the fundamental definition of r becomes

$$(26) \quad r = \frac{\frac{1}{N} \sum_{x,y} (x - \bar{x})(y - \bar{y})f(x, y)}{\sigma_x \sigma_y}.$$

From (22) and (23), we have

$$(x - \bar{x}) = h(u - \bar{u}),$$

and from (24) and (25),

$$(y - \bar{y}) = k(v - \bar{v}).$$

Since (x, y) and (u, v) are merely different notations for the same cell, we have

$$f(x, y) = f(u, v).$$

For computing purposes, the standard deviations are defined as

follows:

$$(27) \quad \begin{aligned} \sigma_x &= h\sigma_u \\ \sigma_y &= k\sigma_v \end{aligned}$$

where

$$(28) \quad \begin{cases} \sigma_u = \left[\frac{1}{N} \sum_u u^2 f(u) - \bar{u}^2 \right]^{1/2} \\ \sigma_v = \left[\frac{1}{N} \sum_v v^2 f(v) - \bar{v}^2 \right]^{1/2} \end{cases}$$

Therefore, (26) becomes

$$r = \frac{\frac{1}{N} \sum_{u,v} (u - \bar{u})(v - \bar{v})f(u, v)}{\sigma_u \sigma_v}.$$

If now we let

$$U = \sum_u u f(u, v) \quad \text{and} \quad V = \sum_v v f(u, v),$$

then since

$$\sum_{u,v} u v f(u, v) = \sum_v v \sum_u u f(u, v) = \sum_u u \sum_v v f(u, v),$$

the above expression for r may be written in either of the following ways:

$$(29) \quad \begin{aligned} r &= \frac{\frac{1}{N} \sum_v v U - \bar{u} \bar{v}}{\sigma_u \sigma_v} \\ &= \frac{\frac{1}{N} \sum_u u V - \bar{u} \bar{v}}{\sigma_u \sigma_v}. \end{aligned}$$

The fact that

$$\sum_v v U = \sum_u u V$$

serves as a check in the table.

The above procedure is illustrated in Table 35.

Explanation: The table is self-explanatory except possibly the U and V entries. Recalling that $U = \sum_u u f(u, v)$, the first entry in the U column is obtained from the sum of the following products: $0 \cdot 1 + 1 \cdot 2 + 2 \cdot 3 + 3 \cdot 1 = 11$; the second entry from $-1 \cdot 1 + 0 \cdot 3 + 1 \cdot 8 + 2 \cdot 1 + 3 \cdot 5 = 24$. Since $V = \sum_v v f(u, v)$ the first

TABLE 35 — COMPUTATION OF r FOR DATA OF TABLE 32

	u	-3	-2	-1	0	1	2	3	$f(v)$	$vf(v)$	$v^2f(v)$	U	vU
v	y	67	72	77	82	87	92	97					
3	92				1	2	3	1	7	21	63	11	33
2	87			1	3	8	1	5	18	36	72	24	48
1	82	4	4	6	4	9	1		28	28	28	-15	-15
0	77	3	3	7	6	4			23	0	0	-18	0
-1	72	2	3	5	6	1	1		18	-18	18	-14	14
-2	67	3	2						5	-10	20	-13	26
-3	62	1							1	-3	9	-3	9
$f(u)$		13	12	19	20	24	6	6	100	54	210		(115)
$uf(u)$		-39	-24	-19	0	24	12	18	-28				
$u^2f(u)$		117	48	19	0	24	24	54	286				
V		-7	-3	3	7	30	11	13					
uV		21	6	-3	0	30	22	39	(115)				

Check

entry in the V row is obtained from $1.4 + 0.3 + -1.2 + -2.3 + -3.1 = -7$. Similarly, for the other entries.

Computations:

$$\sigma_u^2 = \frac{1}{N} \sum u^2 f(u) - \bar{u}^2 = 2.86 - (-.28)^2$$

$$= 2.7816.$$

$$\sigma_u = \sqrt{2.7816} = 1.67.$$

$$\sigma_v^2 = \frac{1}{N} \sum v^2 f(v) - \bar{v}^2 = 2.10 - (.54)^2$$

$$= 1.8084.$$

$$\sigma_v = \sqrt{1.8084} = 1.34.$$

Therefore from (29) we have

$$r = \frac{1.15 - (-.28)(.54)}{(1.67)(1.34)} = 0.58.$$

14. Remarks on Computation of r . (a) *Sign of r .* It should be observed that the sign of r depends on the choice of the positive direction along each coördinate axis. In Table 35 the origin of reference is chosen so that the data occur in the first quadrant and the directions on the (x, y) -axes are the conventional ones. These directions were preserved in changing to (u, v) coördinates. If we had reversed the direction of the v -axis by labeling the y values larger than $y = 77$ by $v = -1, -2, -3$, and those less than $y = 77$ by $v = 1, 2, 3$, the sign of r would be changed. But if the directions of both u and v were reversed, the sign of r would be unchanged.

(b) *Grouping errors.* When N is small, say less than 100, and the data are grouped into cells, grouping errors are introduced. In general, the fewer cells used, the greater the errors. These may be corrected, in part, by applying Sheppard's corrections to σ_u and σ_v . However, this will not be insisted upon in this course.

(c) *Commercial charts.* Computations can be expedited by the use of commercially prepared correlation charts. Several types of chart are available on the market. In her book (reference 15), Professor Helen M. Walker explains the merits of two of these which are recommended. She also gives the following advice to beginners: "A chart is not a crutch to help the novice. It is a means of speeding up operations after they are well understood."

Exercises

1. By equation (29), show that r is independent of the choice of origin and of the units of measurement.
2. In Table 35, evaluate the following sums:

$$\sum_u f(u, 2), \sum_v f(2, v), \sum_u uf(u, 1), \sum_v vf(-2, v), \sum_u \sum_v uvf(u, v), \sum_{u,v} uvf(u, v) \\ \frac{1}{f(v)} \sum_u f(u, v) \quad \text{if } v = 0.$$

3. Derive (29).
4. For the table on page 200, find r and \bar{x} , \bar{y} , σ_x , σ_y . Note that x_0 , y_0 , h and k , do not need to be determined to compute r , but are required for the means and standard deviation of x and y .

15. Regression Lines for a Correlation Table. The data of a correlation table may be thought of as dots lying many deep at the centers of the several cells. There are, of course, $f(x, y)$ of these in any cell whose coördinates are (x, y) , and $f(x)$ is the total number of dots in a vertical column whose coördinate is x . Suppose now we

HEIGHTS AND WEIGHTS OF 200 FRESHMEN
(Heights to Nearest $\frac{1}{4}$ Inch; Weights to Nearest $\frac{1}{2}$ Pound)

$\begin{matrix} x \\ y \end{matrix}$	90- 99.5	100-	110-	120-	130-	140-	150-	160-	170-	180-	190-	200- 209.5	$f(y)$
76- 77.9				1									1
74-							1	1	1	1			4
72-				1	1	1	4		1				8
70-			1	2	6	7	6	2	1	2	1	1	29
68-			2	8	17	8	9	2	1	1	1		49
66-			8	16	14	13	6	2	1			1	61
64-		3	8	7	7	3	3	1	1				33
62-	1	4	1	7	1								14
60-													0
58- 59.9		1											1
$f(x)$	1	8	20	42	46	32	29	8	6	4	2	2	200

Ans. $\bar{x} = 138.45$ lbs.; $\bar{y} = 67.82$ in.

$\sigma_x = 19.6$ lbs.; $\sigma_y = 2.8$ in.

$r = 0.48$.

replace all the data in each column by an equal number of data concentrated at the mean of that column. If we denote the ordinate of this mean point by \bar{y}_x , we have

$$(30) \quad \bar{y}_x = \frac{1}{f(x)} \sum_y y f(x, y).$$

Hence, $\bar{y}_x f(x)$ represents the totality of all the values in a column.

For each of the columns there will be a value of (30). Taking the hypothesis that the mean points of the several columns lie approximately on a straight line $\bar{y}_x = m_1 x + k$, we may find m_1 and k under a least-squares criterion of approximation. If, in applying the criterion,

Sec. 15 Regression Lines for a Correlation Table 201

the square of the difference between the observed mean, \bar{y}_x , and the computed mean, \bar{y}_x , for each array, *viz.*, $(\bar{y}_x - m_1x - k)^2$, is weighted with the number $f(x)$ in the array, it turns out that we get the same values for m_1 and k which we obtained when we fitted the regression line of y on x to the scatter diagram.

In proving this, the student of calculus¹ would have an easy task in obtaining the normal equations:

$$(31) \quad \begin{cases} \sum_x (\bar{y}_x - m_1x - k)f(x) = 0 \\ \sum_x (\bar{y}_x - m_1x - k)xf(x) = 0 \end{cases}$$

whose simultaneous solution yields the desired values of m_1 and k . Expanding (31), we have

$$(32) \quad \begin{cases} \sum_x \bar{y}_x f(x) - m_1 \sum_x x f(x) - k \sum_x f(x) = 0 \\ \sum_x \bar{y}_x x f(x) - m_1 \sum_x x^2 f(x) - k \sum_x x f(x) = 0. \end{cases}$$

Since

$$\sum_x \bar{y}_x f(x) = \sum_x \sum_y y f(x, y) = N\bar{y},$$

and

$$\sum_x \bar{y}_x x f(x) = \sum_x x \sum_y y f(x, y) = \sum_{x, y} xy f(x, y),$$

equation (32) becomes

$$(33) \quad \begin{cases} N\bar{y} - m_1 N\bar{x} - Nk = 0 \\ \sum_{x, y} xy f(x, y) - m_1 \sum_x x^2 f(x) - k N\bar{x} = 0. \end{cases}$$

Solving (33) for m_1 and k we find

$$k = \bar{y} - m_1 \bar{x}$$

$$m_1 = \frac{\sum_{x, y} xy f(x, y) - N\bar{x}\bar{y}}{\sum_x x^2 f(x) - N\bar{x}^2} = \frac{r\sigma_y}{\sigma_x},$$

¹ Differentiating partially $\sum f(x)(\bar{y}_x - m_1x - k)^2$ with respect to m and k respectively, and setting the results equal to zero, yields equation (31). Instead of differentiating this expression one may expand it, regard the result as a quadratic in both m and k , and use the theorem of §3, Chapter VII, to obtain (31).

and the equation of our line becomes

$$\bar{y}_x = r \frac{\sigma_y}{\sigma_x} x + \bar{y} - \frac{\sigma_y}{\sigma_x} r \bar{x},$$

that is

$$(8a) \quad \bar{y}_x - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}).$$

Therefore, the best-fitting line for the means of the columns properly weighted, and the best-fitting line for all the dots are one and the same straight line. But from the point of view of a correlation table, a regression line is to be regarded as the equation from which may be estimated the *average* of all the *y*'s associated with a particular value of *x*. In other words, a prediction in the latter case professes to give only the mean result (Figure 38).

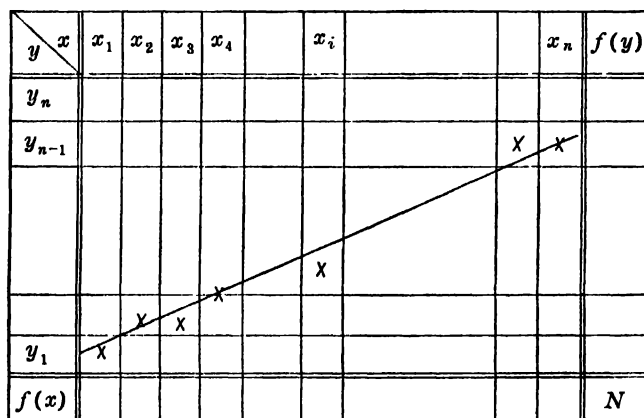


FIG. 38 — THE LINE OF REGRESSION OF *y* ON *x* IS THE BEST FITTING LINE FOR THE MEANS OF THE COLUMNS

16. Applications. The data of a correlation table are usually regarded as a sample of the much larger class of similar data constituting the universe. A regression equation calculated from a limited but representative sample may give valuable estimates of the average values of *y* in the universe associated with designated values of *x*.

Let us consider the data of Table 36 on page 203. Suppose a personnel manager in charge of hiring employees of a manufacturing plant has instituted a system of mental tests for applicants, and has gathered these data showing the relationship between the standing

made by applicants on their mental tests and their productive ability when measured according to a certain standard of production after they are hired.

TABLE 36

x	22.5	27.5	32.5	37.5	42.5	47.5	52.5	57.5	$f(y)$ $=$ $f(v)$	\bar{x}_y
y v	u	-4	-3	-2	-1	0	1	2	3	
125	4					2	3	2		7 47.5
115	3			1	3	1	4	4	4	17 48.1
105	2			5	7	8	11	8	7	46 45.9
95	1		2	1	10	12	9	8	2	44 44.0
85	0	1	3	12	11	7	12	7	1	54 40.7
75	-1	2	1	5	6	16	8	5		43 41.6
65	-2	2	5	5	8	8	6	1		35 38.0
55	-3	2	3	3	4	1	1			14 33.2
$f(x)=f(v)$		7	14	32	49	55	54	35	14	260
\bar{y}_x		67.9	72.1	81.9	84.8	85.7	90.9	95.6	105.0	

Here x represents the grade made on mental test, and y the per cent of standard in production. (See also Table 27.) The means of columns are denoted by \bar{y}_x , and the means of rows \bar{x}_y .

In order to demonstrate to the company's management the connection between his mental tests and the productivity of the employees he has hired, the personnel manager does the following: (1) Computes the coefficient of correlation between the two series; (2) Shows what the estimated productivity of employees would be whose grades in the mental test fell on the mid-points of the class intervals of the mental test data.

The means of the columns and of the rows are given in the table. In addition, he obtains the following results:

$$\bar{x} = 42.17, \quad \sigma_v = 17.41, \quad r = .417,$$

$$\bar{y} = 87.31, \quad \sigma_x = 8.40, \quad m_1 = r \frac{\sigma_y}{\sigma_x} = .864.$$

Therefore, the line of regression of y on x is

$$\bar{y}_x - 87.31 = .864(x - 42.17)$$

or

$$(34) \quad \bar{y}_x = .864x + 50.88.$$

This is the equation of the line that best fits the points which designate the means of the columns (Figure 39). Hence, for an assigned value of x , equation (34) gives the value of y which is the expected mean of the column defined by the assigned value of x . The personnel manager is thus prepared to predict the productivity of applicants on the basis of their mental test grades. In other words, the regression equation calculated from the records of those already hired may be used in selecting from future applicants those most likely to succeed.¹

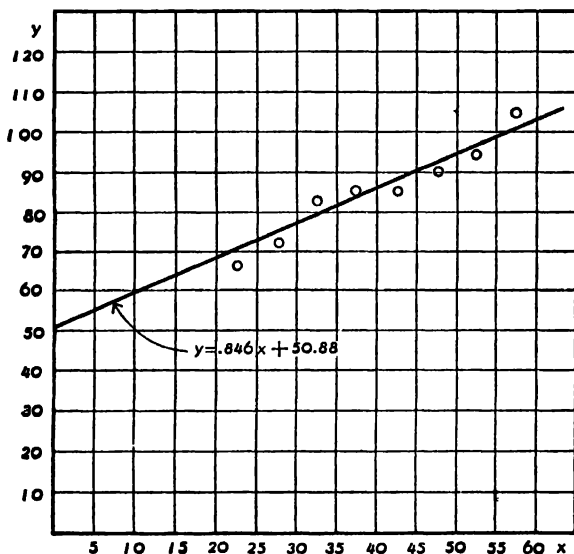


FIG. 39 — MEANS OF COLUMNS AND LINE OF REGRESSION OF y ON x FOR TABLE 36

Exercises

1. Verify the value of r given for Table 36.
2. Verify the means of the columns given in Table 36.
3. Using equation (34) show what the estimated productivity of employees in the factory referred to above would be whose mental test grades were 22.5, 27.5, etc.
4. For Table 35,
 - (a) Find the equations of the regression lines.

¹ The critical reader may doubt if the value $r = .417$ is sufficiently large to warrant much confidence in (34) as a predicting equation. The question of reliability of predictions is discussed later.

(b) Locate the axes through the mean of the table and graph the regression lines.

(c) Compute S_y .

5. As in Exercise 4 for the table on page 200.

Ans. to (a),

$$\bar{y}_x = .069x + 58.3$$

$$\bar{x}_y = 3.36y - 89.4.$$

17. S_y for a Correlation Table. For ungrouped data we have defined S_y as a measure of the clustering of the data around the regression line, and have observed that it is called the standard error of estimate. In order to understand what S_y has to do with "estimates" it is necessary first to consider its meaning in a correlation table. Let us denote by $s_{y \cdot x}$ the standard error about the regression line in the array of y 's at x . Thus we have

$$(35) \quad s_{y \cdot x}^2 = \frac{1}{f(x)} \sum_y (oy - c\bar{y}_x)^2 f(x, y)$$

where oy denotes an observed y value and $c\bar{y}_x$ denotes the value obtained from the regression line for that column. Thus, for the column headed 32.5 in Table 36 we obtain the computed value \bar{y}_x by substituting $x = 32.5$ in (34) whence we find $\bar{y}_x = 78.96$. To evaluate $s_{y \cdot x}^2$ for this column we find the square of the deviation of each of the 32 values of oy from 78.96, add the results and divide by 32. Extracting the square root of the result we find $s_{y \cdot x} = 15.96$. Moving along the regression line suppose we have computed an $s_{y \cdot x}^2$ for each array of y 's and averaged the results. It is interesting to learn that this average is S_y^2 . This is stated more precisely in the following theorem.

Theorem V. *The arithmetic mean of the values of $s_{y \cdot x}^2$ for the several columns when each $s_{y \cdot x}^2$ is weighted with the frequency in that column is $S_y^2 = \sigma_y^2(1 - r^2)$.*

Proof: Using (35) we have

$$\frac{1}{N} \sum_x f(x) s_{y \cdot x}^2 = \frac{1}{N} \sum_x \sum_y (oy - c\bar{y}_x)^2 f(x, y).$$

Substituting the value given by (8a), §15, in the right member of the above identity we have

$$\frac{1}{N} \sum_x \sum_y \left[y - \left\{ \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \right\} \right]^2 f(x, y),$$

that is

$$\frac{1}{N} \sum_{x,y} \left\{ (y - \bar{y}) - r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \right\}^2 f(x, y)$$

which reduces to $\sigma_y^2(1 - r^2)$. It is left as an exercise for the student to show this.

For Table 36 we find $S_y = 15.83$. In Figure 40 the parallel lines

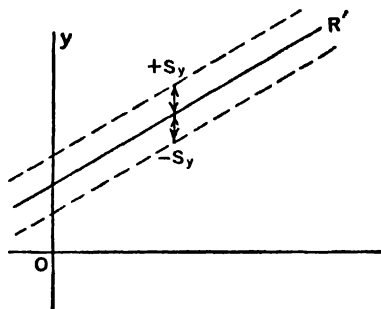


FIG. 40

on either side of the regression line RR' are drawn at a vertical distance of $\pm S_y$ from it. They describe the average limits of scatter above and below the regression line.

To connect S_y with the reliability of predictions it is necessary to introduce the concept of a correlation surface. Indeed, a knowledge of the fundamental properties of a correlation surface

is desirable for a wider outlook on correlation theory in general.

18. Normal Correlation Surface. A correlation table may be idealized into a surface in somewhat the same way that a histogram is idealized into a frequency curve. The concept of a surface relates to the universe from which the observed data of the table may be regarded as a sample. Let the dimensions of the cells of a table be Δx and Δy , and suppose columns are erected upon these cells with altitudes proportional to the frequencies in the cells. The result is a sort of solid histogram. Then as $\Delta x \rightarrow 0$, $\Delta y \rightarrow 0$, $N \rightarrow \infty$, the tops of the columns approach as a limit a smooth surface which is called a correlation surface. Our discussion will be confined to the case where we may assume that this limit is a normal correlation surface. In discussing this surface it is convenient to let x and y represent deviations from the respective means and to let $z = f(x, y)$ denote the frequency function representing the surface. Such a surface is shown in Figure 41.

Any section of this surface parallel to the yz -plane is a normal curve and represents the distribution in a column at x . Similarly any section parallel to the xz -plane representing a row is a normal curve. The frequency in a cell is measured by that portion of the volume under the surface which lies over that cell. All those cells

in which the frequency is a fixed value lie on an ellipse. That is, if contour lines are drawn on the surface joining the points of equal height above the base they will be ellipses. In other words, sections of the surface parallel to the xy -plane are ellipses.

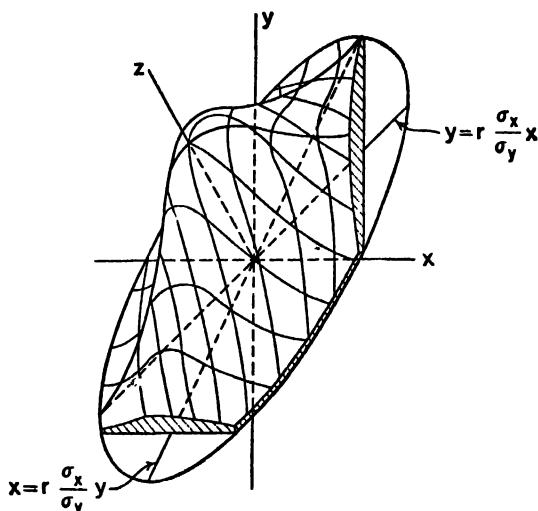


FIG. 41 — FREQUENCY SURFACE FOR CORRELATED VARIABLES

We will digress here for a brief discussion of an ellipse. We may think of an ellipse as a transitional figure between a circle and a straight line, as the circle flattens out. That is to say, the limiting form of an ellipse is a circle at one extreme of the flattening process and a straight line segment at the other extreme. The degree of flatness is called the *eccentricity* of the ellipse, and it is proved in analytic geometry that the eccentricity varies from zero in the case of a circle to unity when the ellipse degenerates into a line. All ellipses having the same eccentricity whatever their size have the same relative proportions and are therefore similar in form.

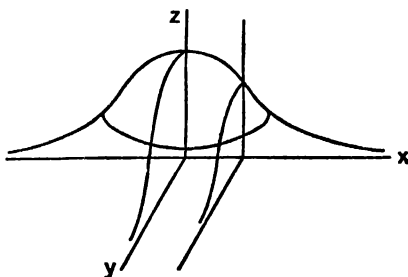


FIG. 42

The eccentricity of the elliptical contours of different normal cor-

relation surfaces varies with the amount of correlation existing in the corresponding universe. A surface with narrow elliptical contours represents a universe in which there is high correlation, whereas if the variables are completely independent in the probability sense the contour lines are circles when the variables are expressed in standard units. If the variables are not expressed in standard units (and $r = 0$) then the contour lines may be ellipses but their major and minor axes will coincide with the x - and y -axes as in Figure 42. When $r \neq 0$ the axes of the ellipses make an angle with the xy -axes, their major axis cuts quadrants I and III in the xy -plane if $r > 0$ (as in Figure 41) and quadrants II and IV if $r < 0$.

19. Properties of Normal Bivariate Surface. The equation of a normal correlation surface is given by

$$(36) \quad f(x, y) = Ke^{-P}$$

where

$$P = \frac{1}{2(1 - r^2)} \left\{ \frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2} \right\},$$

$K = N \div (2\pi\sigma_x\sigma_y\sqrt{1 - r^2})$, and x and y represent the correlated variables referred to their respective means as origin.

By means of (36) an observed distribution may be fitted with the appropriate normal surface assuming that the sample might reasonably have come from such a universe. This is accomplished by replacing σ_x , σ_y , r , and N in (36) by the corresponding statistics calculated from the sample and taking the origin at the mean of the table. Let us assume that an observed distribution has been graduated by such a surface and the theoretical cell frequencies obtained. The surface extends to infinity in the xy -plane but contour ellipses can be obtained which will enclose any desired percentage of the given frequency when these ellipses are projected orthogonally onto the xy -plane. They are all concentric, similar, and similarly placed. Figure 43 represents such an ellipse, say the smallest one necessary to enclose all the given cells. The systems of perpendicular chords represent the columns and rows of the table.

The graduated frequencies for each column are normal distributions whose means lie on the regression line of y on x and whose standard deviations are in each case given by $S_y = \sigma_y(1 - r^2)^{1/2}$. To state the same thing in a slightly different way, an array of y 's corresponding to a fixed value x_1 of x is a normal distribution whose

mean deviates from \bar{y} by $r(\sigma_y/\sigma_x)x_1$ and whose standard deviation is $S_y = \sigma_y(1 - r^2)^{1/2}$ which is independent of x_1 and therefore is the same for all such arrays. Similarly an array of x 's corresponding to a particular value y_1 of y is a normal distribution with a mean which deviates from \bar{x} by $r(\sigma_x/\sigma_y)y_1$, and a standard deviation of $S_x = \sigma_x(1 - r^2)^{1/2}$ which is independent of y_1 and therefore is the same for all such arrays. A careful study of Figure 41 will help in understanding what is meant by these statements.

When the means \bar{y}_x of the columns fall exactly on the regression line, $s_{y \cdot x}$ becomes the standard deviation of a column and is therefore the same as S_y . Theorem V states that S_y^2 is an average of the values of $s_{y \cdot x}^2$ but when all the quantities being averaged have the same value, as they do in the ideal case of the normal surface, their (mean) average is that value. When the standard deviations of the columns are equal, the regression system of y on x is called a *homoscedastic* system. In a universe where they are not equal the system is said to be *heteroscedastic*. For a homoscedastic system with linear regression, $S_y = \sigma_y(1 - r^2)^{1/2}$ is the standard deviation of each array of y 's.

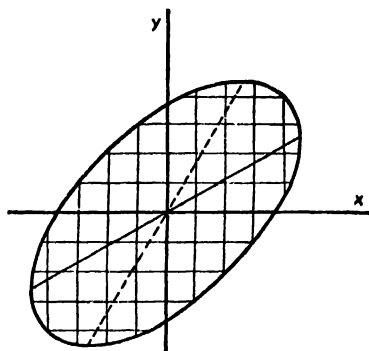


FIG. 43

20. Reliability of Predictions. In using a regression equation to make predictions we are naturally interested in the degree of confidence to be expected in the predictions thus made. The use of S_y in this connection is based upon the properties of the normal correlation surface.

Let us imagine the universe of which Table 36 is a sample and assume that it may be described by a normal surface. Confining our attention to a section parallel to the yz -plane in Figure 41 we know that an x array of y 's is distributed normally about a value of y determined by a designated value of x in the regression equation of y on x . That is, the mean of this normal distribution is the predicted value of y and its standard deviation is S_y . The percentage distribution of such an array is the same as that given in Figure 23 of Chapter VI, if S_y is taken as the unit of measurement along the horizontal axis. But an estimate of S_y is its value cal-

culated from the sample. Moreover, for an observed distribution, we have seen that S_y is the average standard deviation of the several columns and therefore it may reasonably be taken as an approximation to the theoretical S_y which in the universe is the same for all the columns. We also take the calculated regression equation as an approximation to the theoretical.

By measuring deviations from the predicted value in terms of S_y in the same way that σ is used as a unit in measuring deviations from the mean, we may then enter a normal probability scale for

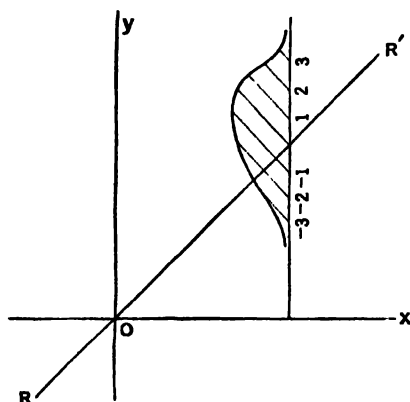


FIG. 44 — REPRESENTING AN x ARRAY OF y 's AND DEVIATIONS OF $\pm S_y$ FROM A PREDICTED VALUE OF y

the probability of a deviation involving multiples of S_y . According to this scale the probability P_y is about .68 for a deviation of $\pm S_y$ from the predicted value, and the chances are even for a deviation of $.6745 S_y$ on either side of the predicted value.

For Table 36 we have found $S_y = 15.83$ and for an applicant making $x = 32.5$ on the mental test we have predicted $y = 78.96$. Therefore the chances are about 68 in 100 that his percentage of productivity will be between $78.96 - 15.83$ and

$78.96 + 15.83$, that is, between 63.13 and 94.79. In other words, the probability is about .68 that the predicted value will not be in error by more than 15.83.

To summarize, in a normal bivariate universe each array is a normal distribution and therefore its mean coincides with its mode. Since regression is linear, a value predicted from the regression equation of y on x is the mean value of y for a designated value of x .

Then, $P_y = \int_{-t}^t \phi(t) dt$ is the probability for a deviation from the predicted value of \bar{y}_x as small as $|t|$ where t is expressed in units of the standard error S_y of a column. Thus,

$$t = \frac{y - \bar{y}_x}{S_y}.$$

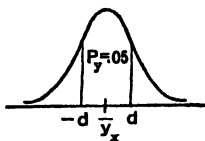
Then $1 - P_y$ is the probability for a deviation as large as $|t|$. Similarly, when dealing with the regression line of x on y , $P_x = \int_{-t}^t$ is the probability for a deviation from the predicted value \bar{x}_y as small as $|t|$, where now $t = (x - \bar{x}_y)/S_x$.

Exercises

1. Refer to problem 4, §4. Assume that the data given there are obtained from a correlation table which is a representative sample from a normal bivariate universe describing the heights and weights of senior men students in colleges and universities of the United States. Then a value predicted from the regression equation of y on x will give the mean of the "column" at x . Similarly, for an assigned y , the corresponding x in the regression equation of x on y will be the mean of the "row" at y . Under this assumption, determine the probability that Doe's height is outside the interval 65.75 - 77.75 inches. What are the chances that Roe will be between 100.8 and 122.4 pounds in weight?
Ans. $1 - P_y = .0027$, $P_x = .5$ (approximately).
2. Discuss the reliability of the predictions which you made in Exercise 3, §16.
Outline of Solution. Suppose a reliability level of $P_y = .5$ is desired. Making the necessary assumptions, this allows a deviation of $t = \pm .6745$. Since $S_y = 15.83$ we have

$$\pm t = \frac{d}{15.83}$$

where $d = y - \bar{y}_x$. That is, $y = \bar{y}_x \pm ?$ For $x = 37.5$, $y = ? \pm ?$ So the probability is .5 that the standard of production will be between what limits for a person making $x = 37.5$ on the mental test? The problem is analogous for any other designated value of P_y and for other assigned values of x .



3. Consider the surface represented by (36). Prove that a section of the surface parallel to the yz coordinate plane is a normal curve with its mean on the regression line of y on x and with variance $S_y^2 = \sigma_y^2(1 - r^2)$.
Outline of Solution. Write (36) in the form

$$(a) \quad f = Ke^{-P},$$

where $P = (u^2 - 2ruv + v^2)/2(1 - r^2)$, $u = x/\sigma_x$, $v = y/\sigma_y$, $\bar{x} = 0 = \bar{y}$. The trace of the surface in the plane $u = u_1$ is determined by substituting u_1 for u in (a). This substitution yields the result

$$(b) \quad f = Ce^{-T}$$

where $T = (v - ru_1)^2/2(1 - r^2)$, $C = Ke^{-u_1^2/2}$. Upon returning to (x, y) coördinates, (b) becomes

$$(c) \quad f = Ce^{-h^2(y-m)^2},$$

where $m = rx_1\sigma_y/\sigma_x$, $h^2 = 1/(2S_y^2)$, $S_y^2 = \sigma_y^2(1 - r^2)$.

21. Non-Linear Regression. Correlation Ratio. We have seen that the regression systems of a normal correlation surface are linear. In a correlation table which is a representative sample from a normal bivariate universe the means of the arrays would lie approximately on straight lines. But in correlation tables which are samples of other types of universes, regression might not be linear. Moreover, one of the regression curves might be strictly linear and the other non-linear. The following numerical example illustrates the latter possibility.

$\begin{array}{c} x \\ y \end{array}$	0	1	2	3	$f(y)$
2	0	0	0	2	2
1	0	1	1	2	4
0	1	1	0	0	2
$f(x)$	1	2	1	4	8

In this example, the regression of y on x is linear whereas that of x on y is non-linear.

When the means of the columns (or of the rows) do not lie approximately on a straight line, the use of r may be misleading because $r = 0$ indicates absence of linear correlation only and not necessarily absence of correlation in general.

One of the best treatments of this situation is that given in the *Carus Monograph on Mathematical Statistics*, which will be reproduced substantially here.

In introducing a *correlation ratio*, η_{yx} , (η ta) of y on x , as an appropriate measure of correlation to take the place of the correlation coefficient in such a situation, we may get suggestions as to what is appropriate by solving for r in (10). This gives

$$(37) \quad r^2 = 1 - \frac{S_y^2}{\sigma_y^2},$$

where we may recall that S_y^2 is the mean square of deviations from the line of

regression. Then

$$r = \pm \left[1 - \frac{S_y^2}{\sigma_y^2} \right]^{1/2}.$$

This formula could be used appropriately as a definition of r in place of our definition in (1), and its examination may throw further light on the significance of r . When $S_y = 0$, the formula gives $r = \pm 1$ and, as we have seen earlier, all the dots of the scatter diagram must then fall exactly on the line of regression. When $S_y = \sigma_y$, the formula gives $r = 0$, and the regression line is in this case of no aid in predicting the value of y from assigned values of x . In the formula $r^2 = 1 - S_y^2/\sigma_y^2$ it is important to keep in mind that the mean square deviation S_y^2 is from the line of regression. Next, let $S_y'^2$ be the corresponding mean square of deviations from the means of columns. Then $S_y'^2 = S_y^2$ when the regression is strictly linear, but $S_y'^2 \neq S_y^2$ when the regression is non-linear. This fact suggests the use of a formula closely related to $[1 - S_y^2/\sigma_y^2]^{1/2}$ for a measure of non-linear regression by replacing S_y by S_y' . We then write

$$(38) \quad \eta_{yx}^2 = 1 - \frac{S_y'^2}{\sigma_y^2}$$

where η_{yx} is the correlation ratio of y on x , and $S_y'^2$ is the mean square of deviations from the means of the columns whether these means are near to or far from the line of regression.

In general, we may say that the correlation ratio of y on x is a measure of the clustering of dots about the means of columns.

An analogous discussion for the rows obviously leads to

$$\eta_{xy}^2 = 1 - \frac{S_x'^2}{\sigma_x^2}$$

giving η_{xy}^2 , the square of the correlation ratio of x on y .

That $\eta_{yx}^2 \leq 1$ and that the equality holds only when all the dots in each column are at the mean of the column follows at once from (38).

That $\eta_{yx}^2 \geq r^2$ may be shown by recalling the meanings of S_y^2 in (37) and of $S_y'^2$ in (38). A mean square of deviations in each column is a minimum when the deviations are taken from the mean of the array. Hence, the $S_y'^2$ in (38) must be equal to or less than S_y^2 in (37) for the same data, since the deviations in (37) are measured from the line of regression. Hence, we have shown that

$$1 \geq \eta_{yx}^2 \geq r^2.$$

Moreover, when the regression of y on x is linear, $\eta_{yx}^2 - r^2$ found from the sample differs from zero by an amount not greater than the fluctuations due to random sampling. Hence, $\eta_{yx}^2 - r^2$ becomes a criterion for testing the linearity of the regression of y on x .

For computational purposes, it is desirable to express the correlation ratios in a form involving the standard deviations of the means of arrays. For this purpose, let \bar{y}_x be the mean of any column of y 's and $\sigma_{\bar{y}_x}$ the standard deviation of the means of columns when the square $(\bar{y}_x - \bar{y})^2$ of each deviation is weighted with the number $f(x)$ in the column. Then it follows that

$$(39) \quad \eta_{yx}^2 = \frac{\sigma_y^2 - S_y'^2}{\sigma_y^2} = \frac{\sigma_{\bar{y}_x}^2}{\sigma_y^2}.$$

That is, the correlation ratio of y on x is the ratio of the standard deviation of the means of columns to the standard deviation of all y 's.¹

To prove (39) we must show that $\sigma_y^2 - S_y'^2 = \sigma_{\bar{y}_x}^2$.^{*} We begin by observing that the concentration of the dots in a column about their mean may be measured in terms of their standard deviation. Let $\sigma_{y \cdot x}$ denote the standard deviation of the y 's in the column at x . That is,

$$(40) \quad \sigma_{y \cdot x}^2 = \frac{1}{f(x)} \sum_y (y - \bar{y}_x)^2 f(x, y).$$

Now, the concentration of the dots in the entire table about the means of the columns may be measured by finding the mean value of all such expressions $\sigma_{y \cdot x}^2$ for all the columns of the table. But since there are more points in some columns than in others, it will be desirable to weight the $\sigma_{y \cdot x}^2$ for each column by multiplying it by the number of points or dots in the column. It is this weighted mean value of the $\sigma_{y \cdot x}^2$'s which we have denoted by $S_y'^2$. That is,

$$(41) \quad S_y'^2 = \frac{1}{N} \sum_x f(x) \sigma_{y \cdot x}^2.$$

In order to verify (39) we must now show that

$$\sigma_y^2 = S_y'^2 + \sigma_{\bar{y}_x}^2.$$

Adapting (14) of §9, Chapter V, to the notation of this chapter, we have

$$(42) \quad N\sigma_y^2 = \sum_x f(x) \sigma_{y \cdot x}^2 + \sum_x f(x) (\bar{y}_x - \bar{y})^2.$$

This follows from the fact that N is composed of the several sub-distributions $f(x)$ in the columns, and $\sigma_{y \cdot x}$ is the standard deviation of a column about its mean \bar{y}_x . It is obvious that

$$\frac{1}{N} \sum_x f(x) (\bar{y}_x - \bar{y})^2$$

gives the variance $\sigma_{\bar{y}_x}^2$ of the means of the columns. The above expression (42) then becomes

$$N\sigma_y^2 = NS_y'^2 + N\sigma_{\bar{y}_x}^2,$$

which reduces to $\sigma_y^2 - S_y'^2 = \sigma_{\bar{y}_x}^2$, and hence we obtain (39).

22. Computation of η^2 . It should be instructive to compute $\eta_{y \cdot x}^2$ for Table 36, by both relations (38) and (39).

¹ Rietz, *Carus Monograph on Mathematical Statistics*, p. 89 *et seq.*

For (38) we have the following:

$$\eta_{yz^2} = 1 - \frac{S_y'^2}{\sigma_y^2}, \quad S_y'^2 = \frac{1}{N} \sum_x f(x) \sigma_{y \cdot x}^2,$$

$$\sigma_{y \cdot x}^2 = \frac{1}{f(x)} \sum_y (y - \bar{y}_x)^2 f(x, y).$$

$\sigma_{y \cdot x}^2$	$f(x)$
106.12 ¹	7
191.83	14
246.48	32
283.63	49
257.65	55
294.51	54
222.53	35
71.43	14

$$S_y'^2 = 246.45$$

$$\sigma_y^2 = (17.41)^2 = 303.11$$

$$\eta_{yz^2} = 1 - \frac{246.45}{303.11}$$

$$= .1869.$$

For (39) we have the following:

$$\eta_{yz^2} = \frac{\sigma_{\bar{y}_x}^2}{\sigma_y^2}, \quad \sigma_{\bar{y}_x}^2 = \frac{1}{N} \sum_x (\bar{y}_x - \bar{y})^2 f(x),$$

$$\bar{y} = 87.31.$$

\bar{y}_x	$f(x)$
67.86	7
72.14	14
81.87	32
84.80	49
85.73	55
90.92	54
95.57	35
105.00	14

$$\sigma_{\bar{y}_x}^2 = 56.66 \text{ (see Exercise 3, p. 97).}$$

$$\eta_{yz^2} = \frac{56.66}{303.11}$$

$$= .1869.$$

¹ See Table 27 and Table 16.

In verifying (39) for this example we have $\sigma_y^2 - S_y'^2 = 303.11 - 246.45 = 56.66$ and $\sigma_{\bar{y}_x}^2 = 56.66$.

The above illustrations are useful in giving an understanding of the meaning of η_{yx}^2 . However, for computational purposes, another formula may be derived which involves less labor than either (38) or (39). In fact, the computation of a correlation ratio may be very conveniently performed by an easy extension of a correlation table. The derivation of the appropriate formula will now be given.

The standard deviation ($\sigma_{\bar{y}_x}$) of the means of the columns may be expressed in the (u, v) units by the relation $\sigma_{\bar{y}_x}^2 = k^2 \sigma_{\bar{v}_u}^2$

where
$$\sigma_{\bar{v}_u}^2 = \frac{1}{N} \sum_u f(u) \bar{v}_u^2 - \bar{v}^2$$

which is the definition of the standard deviation of the variable \bar{v}_u . This is apparent if we observe that the mean for the whole table in the v -direction (\bar{v}) is the mean of the quantities \bar{v}_u for the several columns.¹

Since

$$\bar{v}_u = \frac{1}{f(u)} \sum_v v f(u, v) = \frac{V}{f(u)},$$

we have

$$\sigma_{\bar{v}_u}^2 = \frac{1}{N} \sum_u \frac{V^2}{f(u)} - \bar{v}^2.$$

Recalling that $\sigma_y^2 = k^2 \sigma_v^2$, we have

$$\eta_{yx}^2 = \frac{\sigma_{\bar{y}_x}^2}{\sigma_y^2} = \frac{k^2 \left\{ \frac{1}{N} \sum_u \frac{V^2}{f(u)} - \bar{v}^2 \right\}}{k^2 \sigma_v^2};$$

that is,

$$(43) \quad \eta_{yx}^2 = \frac{1}{\sigma_v^2} \left\{ \frac{1}{N} \sum_u \frac{V^2}{f(u)} - \bar{v}^2 \right\}.$$

An analogous discussion for the rows of x 's leads to

$$(44) \quad \eta_{xy}^2 = \frac{1}{\sigma_u^2} \left\{ \frac{1}{N} \sum_v \frac{U^2}{f(v)} - \bar{u}^2 \right\},$$

giving the square of the correlation ratio of x on y .

$$^1 \frac{1}{N} \sum_u f(u) \bar{v}_u = \frac{1}{N} \sum_u f(u) \frac{1}{f(u)} \sum_v v f(u, v) = \frac{1}{N} \sum_u \sum_v v f(u, v) = \frac{1}{N} \sum_{u,v} v f(u, v) = \bar{v}.$$

Example. Find η_{yz}^2 for Table 35. *Solution.* Referring to this table and using (43) we obtain the following results:

V^2	49	9	9	49	900	121	169	Sum
$V^2/f(u)$	3.78	.75	.47	2.45	37.50	20.17	28.17	93.29

$$\bar{v}^2 = .2916, \quad \sigma_v^2 = 1.8084, \quad N = 100.$$

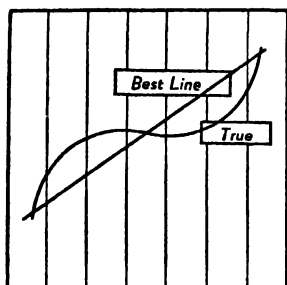
$$\eta_{yz}^2 = \frac{1}{1.8084} \left[\frac{1}{100} (93.29) - .2916 \right]$$

$$\eta_{yz}^2 = .3546.$$

It may be well to mention that the value of η is not independent of the classification of the data. As the class intervals become narrower, η approaches unity. This may be understood from (38). If the grouping were so fine that only one item appeared in each column, then it would constitute the mean of that column. In this case S_v' would be zero and η would therefore be unity. On the other hand, a very coarse grouping tends to make the value of η approach r . "Student" has given a formula for *The Correction to be Made in the Correlation Ratio for Grouping* in *Biometrika*, vol. IX, pp. 316-320.

23. Further Discussion. Test for Linearity of Regression. Let us consider the totality of mean points (x, \bar{y}_x) of the columns and think of a curve connecting them. Of course, for a table of observed data, it is possible to draw many such curves. In order to show clearly why a comparison of η^2 and r^2 is the basis of a test for linearity of regression, it will be necessary to consider a theoretical table in which there is only one such curve. When we speak of *the* regression curve we are thinking, not of the given table in which the dimensions of the cells are h and k , but of an ideal table in which there is an infinity of cells of zero dimensions. To put it another way, consider a sample of N pairs of values (x_i, y_i) from which a correlation table is made with cells whose dimensions are h and k . If parallelepipeds are erected on the cells with heights proportional to the frequencies, the result is a solid histogram bounded by a broken surface. As $h \rightarrow 0$, $k \rightarrow 0$, and $N \rightarrow \infty$, this histogram will approach some solid, bounded by a smooth surface. An example of such a surface is the normal correlation surface. In such an ideal table, it is possible to have but one curve connecting the means of the columns. This curve is sometimes styled the *true* regression curve of y on x . In an

analogous way for the means of the rows there would be a *true* regression curve of x on y . It is one of these curves that we have in mind when we speak of "the regression curve" or "the regression." For a normal bivariate universe (represented by a normal correlation surface), regression is linear. But for other types of bivariate universes (which might be represented by skew surfaces), it is conceivable that regression might be parabolic or exponential or some other type of curve. In such types, regression is said to be



non-linear. The curve which is chosen to approximate the true regression curve must not be confused with the true regression curve. The latter notion relates to the ideal universe from which the data at hand are a sample. It is defined as the locus of the mean points of the columns of the theoretical table. When we fit a curve to the means of the columns of an observed table, this regression curve is merely an

approximation to the ideal set up in the definition. Similar statements may be made about the regression of x on y .

We will now recapitulate the expressions used in the comparative analysis of r^2 and $\eta_{y \cdot x}^2$ for an observed table.

$$(45) \quad \left\{ \begin{array}{l} \sigma_{y \cdot x}^2 = \frac{1}{f(x)} \sum_y (y - \bar{y}_x)^2 f(x, y) \\ S_y'^2 = \frac{1}{N} \sum_x \sigma_{y \cdot x}^2 f(x) \\ \eta_{y \cdot x}^2 = 1 - \frac{S_y'^2}{\sigma_y^2} = \frac{\sigma_{\bar{y}_x}^2}{\sigma_y^2} \end{array} \right.$$

$$(46) \quad \left\{ \begin{array}{l} s_{y \cdot x}^2 = \frac{1}{f(x)} \sum_y (y - \bar{y}_x)^2 f(x, y) \\ S_y^2 = \frac{1}{N} \sum_x s_{y \cdot x}^2 f(x) \\ r^2 = 1 - \frac{S_y^2}{\sigma_y^2} \end{array} \right.$$

Recall that $\sigma_{y \cdot x}^2$ is defined as the variance in a column and therefore as the square of the standard error about the regression curve, what-

ever it may be, which goes through the means of the columns. $S_y'^2$ is an average of the $\sigma_{y \cdot x}^2$ values, and η_{yx}^2 is defined in terms of S_y' . Correspondingly, $s_{y \cdot x}^2$ is the square of the standard error in a column about the *line* which best fits the means of the columns. S_y^2 is an average of the $s_{y \cdot x}^2$ values, and r^2 is defined in terms of S_y . If regression is linear, the means of the columns will fall on the "best-fitting line" and $\sigma_{y \cdot x}^2$ becomes the same as $s_{y \cdot x}^2$. Then $S_y'^2 = S_y^2$, and hence $\eta_{yx}^2 = r^2$.

It is interesting to observe that $\sigma_{y \cdot x}^2$ is the second moment about the mean, for an array of y 's, *i.e.*, for a column. In the notation of moments it could be denoted by $\mu_{2:y \cdot x}$. In this notation, $s_{y \cdot x}^2$ could be denoted by $\nu_{2:y \cdot x}$, being the second moment in an array of y 's about a point other than its mean. Since $\mu_2 \leq \nu_2$, it follows that $\sigma_{y \cdot x}^2 \leq s_{y \cdot x}^2$. Therefore $S_y'^2 \leq S_y^2$ and $\eta_{yx}^2 \geq r^2$. If each y value of a column is at the mean of that column then it is obvious that $\sigma_{y \cdot x}^2$ will be zero. In this case, $S_y' = 0$, and $\eta_{yx}^2 = 1$. On the other hand, for any column, the contribution of $\sigma_{y \cdot x}^2 f(x)$ to $S_y'^2$ cannot exceed its contribution to σ_y^2 . Taking the weighted mean of the respective contributions over all the columns, we have $S_y'^2 \leq \sigma_y^2$ and hence

$$\eta_{yx}^2 \leq 1.$$

Writing (38) in the form

$$S_y' = \sigma_y(1 - \eta_{yx}^2)^{1/2}$$

we see that S_y' is a measure of dispersion about the regression curve (which is the locus of the means) corresponding to $S_y = \sigma_y(1 - r^2)^{1/2}$ which is the standard error about the "best" *line*. If $r^2 = 1$, then y is related to x by a linear function. If $\eta_{yx}^2 = 1$, it follows that y is a single-valued function of x . On the other hand, if $r^2 = 0$, it does not necessarily follow that there is no relation¹ between y and x . If $\eta_{yx}^2 = 0$ then $r^2 = 0$, but if $r^2 = 0$ it does not necessarily follow that $\eta_{yx}^2 = 0$.

In the ideal table, regression of y on x is linear if and only if $\eta_{yx}^2 - r^2 = 0$. But in the case of an observed table, allowance must be made for sampling fluctuations. A corresponding analysis could be made for r^2 and η_{xy}^2 , and $\eta_{xy}^2 - r^2$ computed from the sample should

¹ See H. L. Rietz, "On Functional Relations for which the Coefficient of Correlation is Zero." *Journal American Statistical Association*, vol. 16, 1919, pp. 472-76.

differ from zero by an amount not greater than the fluctuations due to chance, if regression of x on y is linear. The question, naturally arises, what discrepancy between the computed values of η^2 and r^2 may be tolerated before we conclude that regression is non-linear? This problem has been investigated, and Blakeman¹ has proposed a testing formula. If certain assumptions are made, a simple though approximate test may be deduced from Blakeman's formula. According to this approximate test if

$$(47) \quad N(\eta^2 - r^2) < 11.4$$

then linear regression may be assumed. Since there are two η^2 's there are two tests. It is possible for one of the regression curves to be linear and the other not.

Evaluating (47) for Table 35 we obtain $100 [.3546 - (.58)^2] = 1.82$, so the regression of y on x may be assumed to be linear.

R. A. Fisher has shown that the Blakeman test is not very reliable. One can easily construct an example for which regression is obviously non-linear yet which satisfies the criterion (47). Consider the following table:

$y \backslash x$	1	2	3	4	5
3	0	0	1	0	0
2	0	1	0	1	0
1	1	0	0	0	1

Here, $N = 5$, $\sum xy = 27$, $\bar{x} = 3$, $\bar{y} = 9/5$. From (3), therefore, $r = 0$. From (40) and (41), $S_y' = 0$ and $\eta_{yx} = 1$. Applying (47), Blakeman's test yields a verdict of linear regression of y on x . It appears that Blakeman's criterion is of doubtful utility. A more efficient method of testing linearity of regression is given in Part II.

Exercises

1. Using (43) and (44) find η_{yx}^2 and η_{xy}^2 for the table referred to in Exercise 4, page 221. Apply the test (47) and state your opinion about the linearity of regressions.

¹ See *Handbook of Mathematical Statistics*, Rietz and others, p. 131.

2. In the following table, x = Interest Rates, 4-6 months Commercial Paper; y = Total Bills Discounted by Federal Reserve Banks (1923-1932). Find r and η_{yx}^2 . Form an opinion about linearity of regression of y on x . (Data from *Elements of Statistics*, Davis and Nelson, page 288.)

Class Marks y										
7							1	6	6	6
6					1	2	3	4		
5					1	3	1	2		
4				2		9	4	1		
3		1	2	1	4	9	4			
2		1			11	5	1			
1	4		2	3	3	1				
0	2	3	3	5	3					
Class Marks x	0	1	2	3	4	5	6	7	8	9

3. In §44, *Statistical Methods for Research Workers*, R. A. Fisher writes: "The sum of the squares of the deviations of all the values of y from their general mean may be broken up into two parts, one representing the sum of the squares of the deviations of the means of the arrays from the general mean, each multiplied by the number in the array, while the second is the sum of the squares of the deviations of each observation from the mean of the array in which it occurs." [Compare with our (14a) of Chapter V.]

Prove Fisher's statement. *Hint.* In symbols, you are to prove that

$$V = v_1 + v_2$$

where

$$V = \sum_{x,y} (y - \bar{y})^2 f(x, y)$$

$$v_1 = \sum_x (\bar{y}_x - \bar{y})^2 f(x)$$

$$v_2 = \sum_{x,y} (y - \bar{y}_x)^2 f(x, y).$$

4. Prove that η_{yx}^2 is the ratio between v_1 and V as defined in Exercise 3.

5. The mortality experience during the early years of an insurance company presents an interesting study in correlation. The following table shows for male lives the correlation between the ages (x) of the insured at issue of policy and his age (y) at death. Data of the Midland Life Insurance Company,¹ 1906-1924.

$y \backslash x$	15	20	25	30	35	40	45	50	55	60	$f(y)$
70									1	2	3
65								4	9	3	16
60							6	5	7	1	19
55				1		2	12	20	4		39
50					2	13	13	8			36
45				1	12	12	8				33
40			3	13	19	12					47
35		1	8	14	14						37
30		5	10	7							22
25		11	10								21
20	6	4									10
$f(x)$	6	21	31	36	47	39	39	37	21	6	283

Find r , the two η^2 's, and the equations of the lines of regression.

24. Correlation from Ranks. Before defining *rank* we will find the variance of the difference, z , between corresponding values of two variables. Let x and y denote corresponding values of two series each consisting of N variates. Form a third series z where $z_i = x_i - y_i$. Then the mean of z is given by $\bar{z} = \bar{x} - \bar{y}$ and the standard deviation of z is, by definition,

$$\sigma_z^2 = \frac{1}{N} \sum z^2 - \bar{z}^2.$$

¹ From a paper *On Certain Applications of Mathematical Statistics to Actuarial Data* in *The Record*, American Institute of Actuaries, vol. XIII, Part II, No. 28, November, 1924.

Replacing z by its equal $x - y$, we have

$$\begin{aligned}\sigma_z^2 &= \frac{1}{N} \sum (x^2 - 2xy + y^2) - \bar{x}^2 - \bar{y}^2 + 2\bar{x}\bar{y} \\ &= \left\{ \frac{1}{N} \sum x^2 - \bar{x}^2 \right\} - 2 \left\{ \frac{1}{N} \sum xy - \bar{x}\bar{y} \right\} + \left\{ \frac{1}{N} \sum y^2 - \bar{y}^2 \right\}.\end{aligned}$$

Whence

$$(48) \quad \sigma_z^2 = \sigma_x^2 - 2r\sigma_x\sigma_y + \sigma_y^2.$$

If the variables x and y are uncorrelated, we have as a special case

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2.$$

Solving (48) for r , we obtain

$$(49) \quad r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_z^2}{2\sigma_x\sigma_y}.$$

This is another expression for the correlation coefficient and involves standard deviations only. In particular, it may be used to advantage when x and y denote ranks, where by rank we mean order of magnitude or importance. That is, rank refers to the position of a variate in an arrangement.

If x and y denote the ranks of the same item with respect to two characteristics, and no ranks are omitted, and there are no duplications of ranks, then both x and y refer to the integers from 1 to N .

Therefore, $\bar{x} = \bar{y}$, and $\sigma_z^2 = \frac{1}{12} (N^2 - 1) = \sigma_y^2$. See Theorem VI, Chapter V. Moreover,

$$\begin{aligned}\sigma_z^2 &= \frac{1}{N} \sum z^2 - \bar{z}^2 \\ &= \frac{1}{N} \sum (x - y)^2 - (\bar{x} - \bar{y})^2 \\ &= \frac{1}{N} \sum (x - y)^2, \text{ since } \bar{x} - \bar{y} = 0.\end{aligned}$$

Let R denote the correlation coefficient when x and y refer to ranks

rather than variates. Then (49) becomes

$$R = \frac{2\left(\frac{N^2 - 1}{12}\right) - \frac{1}{N} \sum (x - y)^2}{2\left(\frac{N^2 - 1}{12}\right)}$$

which simplifies into

$$(50) \quad R = 1 - \frac{6 \sum (x - y)^2}{N(N^2 - 1)}.$$

This is known as *Spearman's formula* for rank correlation.

If two or more variates are tied it is customary to divide the corresponding rank numbers among the variates concerned, using fractions if necessary.

Example. Suppose we have the following scores made in two tests, arranged in the order of their rank. Find the correlation between ranks.

Individual	1st Subject		2nd Subject		$x - y$	$(x - y)^2$
	Score	Rank = x	Score	Rank = y		
A	92	1	85	2	-1	1
B	86	2	76	4	-2	4
C	84	3	93	1	2	4
D	78	4	68	6	-2	4
E	71	5	67	7	-2	4
F	69	6	83	3	3	9
G	66	7	54	9	-2	4
H	58	8	7Q	5	3	9
I	53	9	43	10	-1	1
J	45	10	59	8	2	4
N = 10					Total	44

We find $R = 1 - \frac{6(44)}{10(99)} = .733.$

Exercises

1. Suppose $z = x + y$. How would this change formulas (48) and (49)?
2. Twelve salesmen are ranked in order of merit for efficiency by their manager. They are also ranked in accordance with their length of service. What indication is there of a relation between length of service and efficiency? (Garrett.)

<i>Salesmen</i>	<i>Years of Service</i>	<i>Order of Merit (Service)</i>	<i>Order of Merit (Effic.)</i>
<i>A</i>	5	7.5	6
<i>B</i>	2	11.5	12
<i>C</i>	10	2	1
<i>D</i>	8	4	9
<i>E</i>	6	6	8
<i>F</i>	4	9	5
<i>G</i>	12	1	2
<i>H</i>	2	11.5	10
<i>I</i>	7	5	3
<i>J</i>	5	7.5	7
<i>K</i>	9	3	4
<i>L</i>	3	10	11

The fractions in the third column denote ties in rank. Thus, *A* and *J* each served 5 years and each is ranked 7.5. The next individual is ranked 9.
Ans. $R = .80$.

3. Find R for the following data:

	<i>Rank</i>	<i>Score</i>	<i>Rank</i>	<i>Score</i>
<i>A</i>	1	92	2	88
<i>B</i>	2	89	4	85
<i>C</i>	3	87	1	93
<i>D</i>	4	86	6	79
<i>E</i>	5	83	7	70
<i>F</i>	6	77	3	87
<i>G</i>	7	71	9	52
<i>H</i>	8	62	5	84
<i>I</i>	9	53	10	41
<i>J</i>	10	40	8	64

Ans. $R = .733$.

25. Interpretation. Common Elements. Although statistical theory gives a description of the indicated relationship between two related variables, the interpretation of the results "abound in pitfalls easily overlooked by the unwary, while they are cantering gaily along upon their arithmetic."

The methodological side has been developed until we can find correlation coefficients by simply turning a crank, but the explanation of the meaning of the result after we find it, needs a brain. . . . No amount of mathematical training and ability can take the place of the judgment and common sense that comes from a knowledge of the field in which the problem lies.¹

In the interpretation of r one should avoid imputing any causal relationship between the variables. In this connection the following pungent remarks of Professor E. B. Wilson² may be appropriately quoted:

Correlation is a mutual affair between two numerical variables; the correlation coefficient r is symmetrical with respect to them. Strictly, y is not correlated with x or x with y , but x and y are correlated. Theory is very important in indicating what facts should be looked for as significant; facts are significant or important largely as they indicate theory, but neither compels the other, as the histories of theorizing and of fact finding amply demonstrate . . . Further, the value of the correlation coefficient depends on the group for which it is determined or on the universe of which that group is a fair sample. The correlation coefficient r of height and weight for a group containing humans from infancy to adult life would be different from, and in fact greater than, the coefficient for college students or for the members of a football squad; there is no such thing as the correlation coefficient per se.

If the student has mastered the underlying mathematical theory he should be able to understand and profit by the interpretations given by the writers in his particular field of interest. As a final aid in forming a conception of its meaning, we state a theorem which gives to r a meaning in pure chance. If x and y are affected by s equally likely causes of which t are common to both, then $r = t/s$.

Theorem VI. *An urn containing white and black balls is so maintained that in drawing a ball the probability of getting a white ball is a constant p and that of getting a black ball is q ($= 1 - p$). The first drawing of a pair of drawings is to consist of s balls taken one at a time from the urn. The second drawing is to consist of s balls of which t are taken at random from the s first drawn, and $s - t$ are drawn one at a time from the urn. Then the correlation coefficient between the numbers of white balls in the two drawings is t/s .*

As an illustration of the theorem we will take $s = 5$, $t = 3$, $p = \frac{1}{2}$. Let x be the number of white balls in the first drawing and y the

¹ Crathorne in Journal of the American Statistical Association, vol. 26, Supplement, March, 1931, p. 27.

² *Correlation and Association*, Journal of the American Statistical Association, vol. 26 (1931), pp. 250-256.

number of white balls in the second drawing. Then Table 37, constructed by the theory of probability,¹ exhibits the *a priori* frequencies when we use as small numbers as possible for frequencies subject to the condition that each frequency is to be an integer.

TABLE 37 — A PRIORI FREQUENCIES

$y \backslash x$	0	1	2	3	4	5	$f(y)$
5	0	0	0	9	6	1	16
4	0	0	81	108	45	6	240
3	0	243	648	432	108	9	1440
2	243	1620	1728	648	81	0	4320
1	1458	3159	1620	243	0	0	6480
0	2187	1458	243	0	0	0	3888
$f(x)$	3888	6480	4320	1440	240	16	16,384

According to the theorem the correlation coefficient should be $\frac{2}{3}$. It is left as an exercise for the student to show, by computing r from the table, that this is actually the case.

Review Questions and Problems

1. Define the following terms: statistics, variate, discrete, class interval, class mark, x -array of y 's, range, regression line, sample, universe, coefficient of variation, variance.
2. Name and define five averages. Discuss their advantages and limitations.
3. What does a ratio chart show that a chart with a uniform scale does not? If you wished to plot data so as to secure the effect of a ratio chart, but had no ratio paper available, how would you accomplish the desired result?
4. Prove the following:
 - (a) The algebraic sum of the deviations of the variates from their mean is zero.
 - (b) The second moment about an arbitrary point equals the second moment about the mean increased by the square of the distance between the arbitrary point and the mean.

¹ Explained in Part II.

5. (a) Define and explain how to compute the following:

$$Q_1, Q_2, Q, MD, s, \sigma.$$

- (b) In the case of a normal distribution give the value of each of the first four constants in (a) in terms of \bar{x} or σ .
6. (a) Give the equation of the normal curve in both arbitrary coördinates and standard units. State the relation between abscissas and between ordinates in the two systems.
- (b) State the properties of the normal curve.
7. Show how to fit a straight line $y = mx + k$ by the method of moments by deriving the expressions for m and k .
8. Show how to fit an exponential function by the method explained in the text.
9. Show how to fit a parabola by the method of moments.
10. (a) Give two of the formulas for r . Discuss the use or uses of correlation in any problem that occurs to you.
- (b) Show that the slope of the line in problem 7 may be written $r\sigma_y/\sigma_x$.
11. Prove that $|r| \leq 1$.
- (b) Define the correlation ratio. Discuss its use.
12. Discuss rank correlation.
13. Derive the following relations:

$$\bar{x} = c\bar{u} + x_0$$

$$\mu_2 = \nu_2 - \nu_1^2$$

$$\mu_{2;x} = c^2\mu_{2;u}$$

$$\sigma_x = c\sigma_u.$$

14. The following is a reduced distribution of the breakfast checks at a cafeteria. Using the indirect method find \bar{x} and σ_x .

x	f
8-12	4
13-17	8
18-22	24
23-27	21
28-32	15
33-37	14
38-42	7
43-47	4
48-52	2
53-57	1

Ans. $\bar{x} = 27.2¢, \sigma = 9.4¢.$

15. Derive the relations which give the third and fourth moments about the mean in terms of moments about an arbitrary origin. Define α_3 and α_4 . What information do they give?
16. Compute the value of α_3 and of α_4 for the distribution in Exercise 14.

17. The following is a distribution of the heights of students where x denotes heights in inches and f is the number of students of the corresponding heights. Find \bar{x} , σ_x , α_3 , and α_4 .

x	f
60.5	1
62.0	3
63.5	14
65.0	32
66.5	61
68.0	80
69.5	71
71.0	35
72.5	24
74.0	2
75.5	1

18. For N values of a variable v it is known that $\sum v = 0$ and $\sum v^2 = N$. What are the origin and unit of v ?
19. Find in two ways the value of P for which the function

$$y = \frac{1}{N} \sum f(x - P)^2$$

has the smallest value.

20. (*Walker*) An algebra test was given to 400 high school children, of whom 150 were boys and 250 were girls. The results were as follows:

$n_1 = 150$	$n_2 = 250$
$\bar{x}_1 = 72.5$	$\bar{x}_2 = 73.6$
$\sigma_1 = 7.0$	$\sigma_2 = 6.4$

Find the mean and standard deviation of the combined groups.

21. For a normal distribution of 1500 students' grades, $\bar{x} = 75$, $\sigma_x = 10$. What values of x will include the middle 500 grades? How many grades were below 60; above 90?
22. Suppose a distribution of 1000 breakfast checks from the cafeteria mentioned in problem 14 showed the following results: $\bar{x} = 27¢$, $\sigma_x = 9¢$, $\alpha_3 = 0$, $\alpha_4 = 3$. On the basis of these results what is the expected frequency in the 23-27¢ class interval?
23. Given the following data as to the heights (y) and weights (x) of college men:

$$\begin{array}{lll} \sum y = 6,800, & \sum y^2 = 463,025, & \sum xy = 1,022,250 \\ \sum x = 15,000, & \sum x^2 = 2,272,500, & N = 100. \end{array}$$

Find \bar{x} , \bar{y} , α_x , σ_y , r .

24. Derive the expression for the standard error of estimate,

$$S_y = \sigma_y(1 - r^2)^{1/2}.$$

25. Discuss the use of S_y in predictions.

26. Compute the median, quartiles, and quartile deviation for the following distribution where x = bushels per acre and f = corresponding frequency.

x	f
1	3
3	26
5	78
7	107
9	113
11	65
13	40
15	22
17	45
19	41
21	21
23	23

27. (a) Find r for the following table using (u, v) coördinates.

$y \backslash x$	17	19	21	23	$f(y)$
18		3	2	1	6
15	2	4	3	1	10
12	2	1	1		4
$f(x)$	4	8	6	2	20

- (b) For the above data, find \bar{x} , \bar{y} , σ_x , σ_y , and the equations of the regression lines.

28. For Table 38, (a) find the correlation coefficient, (b) find the equations of the lines of regression, (c) locate the coördinate axes through the arithmetic mean of the table and plot the lines obtained in (b).
29. Fit an exponential function of the type $y = Ae^{Bx}$ to the following data:

x	0	2	4
y	2	10	100

First find the equation in the forms

(a) $Y = at + b$

(b) $Y = mx + k$

and then determine A and B .

TABLE 38 — CORRELATION TABLE FOR MONTHLY RAINFALL AT IOWA CITY AND DES MOINES, 1890-1925

IOWA CITY

DES MOINES

$y \backslash x$	0.245	0.745	1.245	1.745	2.245	2.745	3.245	3.745	4.245	4.745	5.245	5.745	6.245	6.745	7.245	7.745	8.245	8.745	9.245	9.745	10.245	10.745	$f(y)$
10.245																			1				1
9.745																1							1
9.245																	1	1					2
8.745															1				1				2
8.245													2		1	1							4
7.745											1											1	2
7.245									2	1		1			2	1							7
6.745			2		1		2	1			1				1			1	1				10
6.245											1												1
5.745						4	1		1			1	2				1						10
5.245			1		1	2		2		2	1		1	1					1				12
4.745					2	1	2	2	1	1	1		1			2			1				14
4.245			1		1	1	2		1	4	1	2	1										14
3.745				4	2	1	2	6	5	3	2	2	2		1								30
3.245		2	1	4	6	6	3	7	2		1						1				1		34
2.745			3	4	1	8	4	4	2	1	2		1										30
2.245			1	5	10	7	6	4	2	2													37
1.745	1	4	7	12	13	8	5	1	1	1	2		1										56
1.245	3	8	18	17	6	8	4	2		1													67
0.745	6	16	21	12	6	1	1	1			1			1									66
0.245	13	12	3	4																			32
$f(x)$	23	42	58	62	49	47	32	27	18	15	14	7	10	5	6	5	3	2	5	0	1	1	432

30. How does the scatter diagram assist one in deciding whether the regression is linear or non-linear? Give the formulas for the correlation coefficient and for the correlation ratio of y on x , explaining the meaning of the letters used. How would you use these indices of correlation to decide whether the regression of y on x is linear or non-linear?
31. (a) In a normal distribution in which $\bar{x} = 0$ and $\sigma_x = 4$, what proportion of the data lie where $x > 12$?
- (b) If 100 of the data lie between $x = -6$ and $x = -8$, how many of the data are there in the whole distribution?
32. (a) When the variates are ungrouped what is perhaps the best formula for σ_x ? *Ans.*

$$\sigma_x = \frac{[N \sum x^2 - (\sum x)^2]^{1/2}}{N}.$$

- (b) What does this expression become in terms of N when x refers to the integers from 1 to N ?
33. (a) Expand $(a + b + c + d)^2$.
- (b) The expansion of $(x_1 + x_2 + \dots + x_n)^2$ consists of the sum of the squares of the x 's plus the sum of their products taken two at a time. Express this expansion in summation notation.
34. (a) Show that the formula for MD may be written

$$MD = \frac{2}{N} [\bar{x} \sum_{x_i < \bar{x}} f_i - \sum_{x_i < \bar{x}} f_i x_i].$$

Hint. For $x_i < \bar{x}$, $\sum f_i |x_i - \bar{x}| = -\sum f_i (x_i - \bar{x}) = \sum f_i (\bar{x} - x_i) = \bar{x} \sum f_i - \sum f_i x_i$.

For $x_i > \bar{x}$, $\sum f_i |x_i - \bar{x}| = -\sum f_i (\bar{x} - x_i)$.

Since \bar{x} is the centroid (§14, Chapter III), $-\sum f_i (\bar{x} - x_i)$ for $x_i > \bar{x}$ equals $\sum f_i (\bar{x} - x_i)$ for $x_i < \bar{x}$.

- (b) Using this formula evaluate MD for one of the distributions in the text.
35. Given N pairs of variates: $(x_{11}, x_{21}); (x_{12}, x_{22}); (x_{13}, x_{23}); \dots; (x_{1n}, x_{2n})$. Show that:
- (a) the mean \bar{x} of all the variates is

$$\bar{x} = \frac{1}{2N} \sum_1^n (x_{1i} + x_{2i}),$$

- (b) the variance σ^2 taken about the \bar{x} in (a) is

$$\sigma^2 = \frac{1}{2N} \left(\sum_1^n (x_{1i} - \bar{x})^2 + \sum_1^n (x_{2i} - \bar{x})^2 \right).$$

Note. The quantity

$$r' = \frac{1}{N\sigma^2} \sum_1^n (x_{1i} - \bar{x})(x_{2i} - \bar{x})$$

where \bar{x} and σ^2 are defined as in (a) and (b) is called the *intra-class correlation coefficient*. For its use see *Statistical Methods for Research Workers*, Fisher (§38), Oliver and Boyd, London.

36. Let $S_r = \frac{1}{N} \sum_{x=1}^N x^r$. Prove that $S_1 = N(N+1)/2$,
 $S_2 = N(N+1)(2N+1)/6$, $S_3 = S_1^2$.
37. Sketch the graph of $y = Ae^{Bx}$, $-\infty \leq x \leq \infty$, when (a) both A and B are positive, (b) A is positive and B negative, (c) A is negative and B positive, (d) both A and B are negative.
38. A large number of rectangles are drawn all having the same perimeter but different bases (x) and altitudes (y). Which of the following is the correct answer? The coefficient of correlation between x and y is (a) negative and numerically large, (b) positive and numerically small, (c) positive and numerically large, (d) approximately zero.
39. For N correlated values of x and y the regression equation of y on x is found to be $y = 1 + x$. If $\bar{x} = 0$, $r = 0.5$, and $\sigma_x = 1$, determine \bar{y} and S_y .
40. Let NS_y^2 denote the sum of squares of deviations from the line of least squares (Case I).
- (a) Show that $NS_y^2 = \sum y^2 - m \sum xy - k \sum y$.

$$\begin{aligned} \text{Hint. } NS_y^2 &= \sum (y - mx - k)^2 \\ &= \sum y(y - mx - k) - m \sum x(y - mx - k) \\ &\quad - k \sum (y - mx - k). \end{aligned}$$

The last two expressions vanish. Why?

- (b) If m and k are replaced by their determinant values from (5), p. 143, show that

$$NS_y^2 = \frac{1}{D} \begin{vmatrix} \sum y^2 & \sum y & \sum xy \\ \sum y & N & \sum x \\ \sum xy & \sum x & \sum x^2 \end{vmatrix}, \quad D = \begin{vmatrix} N & \sum x \\ \sum x & \sum x^2 \end{vmatrix}.$$

The third order determinant is D bordered by $\sum y^2$, $\sum y$, $\sum xy$.

- (c) If x and y are replaced by x' and y' , denoting deviations from their respective means, find the values of the resulting determinants in (b).
- (d) From the results in (c) show that $S_y^2 = \sigma_y^2(1 - r^2)$.
41. Discuss the properties of the normal correlation surface and their use in passing judgment on the reliability of predictions based upon the regression line of y on x .
42. (For calculus students) In fitting points in a plane by a line so that the sum of squares of perpendicular deviations shall be a minimum, a second line may be found for which the sum of squares of perpendicular deviations is a maximum. If $\sum d^2$ is the sum of squares of deviations from the first line and $\sum D^2$ is the sum of squares of deviations from the second line, show that $\sum d^2 / \sum D^2 = (1 + r)/(1 - r)$. [Reference: *Bulletin American Mathematical Society*, vol. 47 (1941), p. 710.]

APPENDIX

Tables

- I. ORDINATES AND AREAS OF THE NORMAL CURVE.**
- II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES.**

TABLE I. ORDINATES AND AREAS OF THE NORMAL CURVE, $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$

t	$\phi(t)$	$\int_0^t \phi(t) dt$	t	$\phi(t)$	$\int_0^t \phi(t) dt$	t	$\phi(t)$	$\int_0^t \phi(t) dt$
.00	.39894	.00000	.45	.36053	.17364	.90	.26609	.31594
.01	.39892	.00399	.46	.35889	.17724	.91	.26369	.31859
.02	.39886	.00798	.47	.35723	.18082	.92	.26129	.32121
.03	.39876	.01197	.48	.35553	.18439	.93	.25888	.32381
.04	.39862	.01595	.49	.35381	.18793	.94	.25647	.32639
.05	.39844	.01994	.50	.35207	.19146	.95	.25406	.32894
.06	.39822	.02392	.51	.35029	.19497	.96	.25164	.33147
.07	.39797	.02790	.52	.34849	.19847	.97	.24923	.33398
.08	.39767	.03188	.53	.34667	.20194	.98	.24681	.33646
.09	.39733	.03586	.54	.34482	.20540	.99	.24439	.33891
.10	.39695	.03983	.55	.34294	.20884	1.00	.24197	.34134
.11	.39654	.04380	.56	.34105	.21226	1.01	.23955	.34375
.12	.39608	.04776	.57	.33912	.21566	1.02	.23713	.34614
.13	.39559	.05172	.58	.33718	.21904	1.03	.23471	.34850
.14	.39505	.05567	.59	.33521	.22240	1.04	.23230	.35083
.15	.39448	.05962	.60	.33322	.22575	1.05	.22988	.35314
.16	.39387	.06356	.61	.33121	.22907	1.06	.22747	.35543
.17	.39322	.06749	.62	.32918	.23237	1.07	.22506	.35769
.18	.39253	.07142	.63	.32713	.23565	1.08	.22265	.35993
.19	.39181	.07535	.64	.32506	.23891	1.09	.22025	.36214
.20	.39104	.07926	.65	.32297	.24215	1.10	.21785	.36433
.21	.39024	.08317	.66	.32086	.24537	1.11	.21546	.36650
.22	.38940	.08706	.67	.31874	.24857	1.12	.21307	.36864
.23	.38853	.09095	.68	.31659	.25175	1.13	.21069	.37076
.24	.38762	.09483	.69	.31443	.25490	1.14	.20831	.37286
.25	.38667	.09871	.70	.31225	.25804	1.15	.20594	.37493
.26	.38568	.10257	.71	.31006	.26115	1.16	.20357	.37698
.27	.38466	.10642	.72	.30785	.26424	1.17	.20121	.37900
.28	.38361	.11026	.73	.30563	.26730	1.18	.19886	.38100
.29	.38251	.11409	.74	.30339	.27035	1.19	.19652	.38298
.30	.38139	.11791	.75	.30114	.27337	1.20	.19419	.38493
.31	.38023	.12172	.76	.29887	.27637	1.21	.19186	.38686
.32	.37903	.12552	.77	.29659	.27935	1.22	.18954	.38877
.33	.37780	.12930	.78	.29431	.28230	1.23	.18724	.39065
.34	.37654	.13307	.79	.29200	.28524	1.24	.18494	.39251
.35	.37524	.13683	.80	.28969	.28814	1.25	.18265	.39435
.36	.37391	.14058	.81	.28737	.29103	1.26	.18037	.39617
.37	.37255	.14431	.82	.28504	.29389	1.27	.17810	.39796
.38	.37115	.14803	.83	.28269	.29673	1.28	.17585	.39973
.39	.36973	.15173	.84	.28034	.29955	1.29	.17360	.40147
.40	.36827	.15542	.85	.27798	.30234	1.30	.17137	.40320
.41	.36678	.15910	.86	.27562	.30511	1.31	.16915	.40490
.42	.36526	.16276	.87	.27324	.30785	1.32	.16694	.40658
.43	.36371	.16640	.88	.27086	.31057	1.33	.16474	.40824
.44	.36213	.17003	.89	.26848	.31327	1.34	.16256	.40988

TABLE I. ORDINATES AND AREAS OF THE NORMAL CURVE, $\phi(t)$ $\sqrt{2\pi} e^{-t^2/2}$

t	$\phi(t)$	$\int_0^t \phi(t) dt$	t	$\phi(t)$	$\int_0^t \phi(t) dt$	t	$\phi(t)$	$\int_0^t \phi(t) dt$
1.35	.16038	.41149	1.80	.07895	.46407	2.25	.03174	.48778
1.36	.15822	.41309	1.81	.07754	.46485	2.26	.03103	.48809
1.37	.15608	.41466	1.82	.07614	.46562	2.27	.03034	.48840
1.38	.15395	.41621	1.83	.07477	.46638	2.28	.02965	.48870
1.39	.15183	.41774	1.84	.07341	.46712	2.29	.02898	.48899
1.40	.14973	.41924	1.85	.07206	.46784	2.30	.02833	.48928
1.41	.14764	.42073	1.86	.07074	.46856	2.31	.02768	.48956
1.42	.14556	.42220	1.87	.06943	.46926	2.32	.02705	.48983
1.43	.14350	.42364	1.88	.06814	.46995	2.33	.02643	.49010
1.44	.14146	.42507	1.89	.06687	.47062	2.34	.02582	.49036
1.45	.13943	.42647	1.90	.06562	.47128	2.35	.02522	.49061
1.46	.13742	.42786	1.91	.06439	.47193	2.36	.02463	.49086
1.47	.13542	.42922	1.92	.06316	.47257	2.37	.02406	.49111
1.48	.13344	.43056	1.93	.06195	.47320	2.38	.02349	.49134
1.49	.13147	.43189	1.94	.06077	.47381	2.39	.02294	.49158
1.50	.12952	.43319	1.95	.05959	.47441	2.40	.02239	.49180
1.51	.12758	.43448	1.96	.05844	.47500	2.41	.02186	.49202
1.52	.12566	.43574	1.97	.05730	.47558	2.42	.02134	.49224
1.53	.12376	.43699	1.98	.05618	.47615	2.43	.02083	.49245
1.54	.12188	.43822	1.99	.05508	.47670	2.44	.02033	.49266
1.55	.12001	.43943	2.00	.05399	.47725	2.45	.01984	.49286
1.56	.11816	.44062	2.01	.05292	.47778	2.46	.01936	.49305
1.57	.11632	.44179	2.02	.05186	.47831	2.47	.01889	.49324
1.58	.11450	.44295	2.03	.05082	.47882	2.48	.01842	.49343
1.59	.11270	.44408	2.04	.04980	.47932	2.49	.01797	.49361
1.60	.11092	.44520	2.05	.04879	.47982	2.50	.01753	.49379
1.61	.10915	.44630	2.06	.04780	.48030	2.51	.01709	.49396
1.62	.10741	.44738	2.07	.04682	.48077	2.52	.01667	.49413
1.63	.10567	.44845	2.08	.04586	.48124	2.53	.01625	.49430
1.64	.10396	.44950	2.09	.04491	.48169	2.54	.01585	.49446
1.65	.10226	.45053	2.10	.04398	.48214	2.55	.01545	.49461
1.66	.10059	.45154	2.11	.04307	.48257	2.56	.01506	.49477
1.67	.09893	.45254	2.12	.04217	.48300	2.57	.01468	.49492
1.68	.09728	.45352	2.13	.04128	.48341	2.58	.01431	.49506
1.69	.09566	.45449	2.14	.04041	.48382	2.59	.01394	.49520
1.70	.09405	.45543	2.15	.03955	.48422	2.60	.01358	.49534
1.71	.09246	.45637	2.16	.03871	.48461	2.61	.01323	.49547
1.72	.09089	.45728	2.17	.03788	.48500	2.62	.01289	.49560
1.73	.08933	.45818	2.18	.03706	.48537	2.63	.01256	.49573
1.74	.08780	.45907	2.19	.03626	.48574	2.64	.01223	.49585
1.75	.08628	.45994	2.20	.03547	.48610	2.65	.01191	.49598
1.76	.08478	.46080	2.21	.03470	.48645	2.66	.01160	.49609
1.77	.08329	.46164	2.22	.03394	.48679	2.67	.01130	.49621
1.78	.08183	.46246	2.23	.03319	.48713	2.68	.01100	.49632
1.79	.08038	.46327	2.24	.03246	.48745	2.69	.01071	.49643

TABLE I. ORDINATES AND AREAS OF THE NORMAL CURVE, $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$

t	$\phi(t)$	$\int_0^t \phi(t) dt$	t	$\phi(t)$	$\int_0^t \phi(t) dt$	t	$\phi(t)$	$\int_0^t \phi(t) dt$
2.70	.01042	.49653	3.15	.00279	.49918	3.60	.00061	.49984
2.71	.01014	.49664	3.16	.00271	.49921	3.61	.00059	.49985
2.72	.00987	.49674	3.17	.00262	.49924	3.62	.00057	.49985
2.73	.00961	.49683	3.18	.00254	.49926	3.63	.00055	.49986
2.74	.00935	.49693	3.19	.00246	.49929	3.64	.00053	.49986
2.75	.00909	.49702	3.20	.00238	.49931	3.65	.00051	.49987
2.76	.00885	.49711	3.21	.00231	.49934	3.66	.00049	.49987
2.77	.00861	.49720	3.22	.00224	.49936	3.67	.00047	.49988
2.78	.00837	.49728	3.23	.00216	.49938	3.68	.00046	.49988
2.79	.00814	.49736	3.24	.00210	.49940	3.69	.00044	.49989
2.80	.00792	.49744	3.25	.00203	.49942	3.70	.00042	.49989
2.81	.00770	.49752	3.26	.00196	.49944	3.71	.00041	.49990
2.82	.00748	.49760	3.27	.00190	.49946	3.72	.00039	.49990
2.83	.00727	.49767	3.28	.00184	.49948	3.73	.00038	.49990
2.84	.00707	.49774	3.29	.00178	.49950	3.74	.00037	.49991
2.85	.00687	.49781	3.30	.00172	.49952	3.75	.00035	.49991
2.86	.00668	.49788	3.31	.00167	.49953	3.76	.00034	.49992
2.87	.00649	.49795	3.32	.00161	.49955	3.77	.00033	.49992
2.88	.00631	.49801	3.33	.00156	.49957	3.78	.00031	.49992
2.89	.00613	.49807	3.34	.00151	.49958	3.79	.00030	.49992
2.90	.00595	.49813	3.35	.00146	.49960	3.80	.00029	.49993
2.91	.00578	.49819	3.36	.00141	.49961	3.81	.00028	.49993
2.92	.00562	.49825	3.37	.00136	.49962	3.82	.00027	.49993
2.93	.00545	.49831	3.38	.00132	.49964	3.83	.00026	.49994
2.94	.00530	.49836	3.39	.00127	.49965	3.84	.00025	.49994
2.95	.00514	.49841	3.40	.00123	.49966	3.85	.00024	.49994
2.96	.00499	.49846	3.41	.00119	.49968	3.86	.00023	.49994
2.97	.00485	.49851	3.42	.00115	.49969	3.87	.00022	.49995
2.98	.00471	.49856	3.43	.00111	.49970	3.88	.00021	.49995
2.99	.00457	.49861	3.44	.00107	.49971	3.89	.00021	.49995
3.00	.00443	.49865	3.45	.00104	.49972	3.90	.00020	.49995
3.01	.00430	.49869	3.46	.00100	.49973	3.91	.00019	.49995
3.02	.00417	.49874	3.47	.00097	.49974	3.92	.00018	.49996
3.03	.00405	.49878	3.48	.00094	.49975	3.93	.00018	.49996
3.04	.00393	.49882	3.49	.00090	.49976	3.94	.00017	.49996
3.05	.00381	.49886	3.50	.00087	.49977	3.95	.00016	.49996
3.06	.00370	.49889	3.51	.00084	.49978	3.96	.00016	.49996
3.07	.00358	.49893	3.52	.00081	.49978	3.97	.00015	.49996
3.08	.00348	.49897	3.53	.00079	.49979	3.98	.00014	.49997
3.09	.00337	.49900	3.54	.00076	.49980	3.99	.00014	.49997
3.10	.00327	.49903	3.55	.00073	.49981			
3.11	.00317	.49906	3.56	.00071	.49981			
3.12	.00307	.49910	3.57	.00068	.49982			
3.13	.00298	.49913	3.58	.00066	.49983			
3.14	.00288	.49916	3.59	.00063	.49983			

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

Prop. Parts		N									
		150	638	667	696	725	754	782	811	840	869
20	28	17 898	926	955	98	'01.	'041	'070	'099	'127	'156
2.9	2.8	18 184	21.	241	270	298	327	35.	384	412	441
5.8	5.6	469	498	526	554	583	611	639	66.	696	724
8.7	8.4	18 752	780	808	83.	865	893	921	949	97.	'005
11.6	11.2	19 033	061	089	117	14.	173	201	229	25.	285
14.5	14.0	312	340	368	396	424	451	479	507	535	562
17.4	16.8										
20.3	19.6		590	618	645	673	700	728	756	783	811
23.2	22.4	19 866	893	921	948	976	*003	*030	*058		'112
26.1	25.2	20 140	167	194	222	249	276	303	330		385
		160	412	439	466	49.	520	548	575	602	629
											656
27	28	683	710	737	763	790	817	844	871	898	925
2.7	2.6	20 952	978	*005	'032	*059	'085	'112	'139	'165	'192
5.4	5.2	21 219	245	272	299	325	352	378	405	431	458
8.1	7.8										
10.8	10.4	484	511	537	564	590	617	643	669	696	722
13.5	13.0	21 748	775	801	827	854	880	906	932	958	985
16.2	15.6	22 011	037	063	089	115	141	167	194	220	246
18.9	18.2										
21.6	20.8	272	298	324	350	376	401	427	453	479	505
24.3	23.4	531	557	583	608	634	660	686	712	737	763
		22 789	814	840	866	891	917	943	968	994	*019
		170	23 045	070	096	121	147	172	198	223	249
											274
	25	300	325	350	376	401	426	452	477	502	528
1	2.5	553	578	603	629	654	679	704	729	754	779
2	5.0	23 805	830	855	880	905	930	955	980	'005	'030
3	7.5										
4	10.0	24 055	080	105	130	155	180	204	229	254	279
5	12.5	304	329	353	378	403	428	452	477	502	527
6	15.0	551	576	601	625	650	674	699	724	748	773
7	17.5										
8	20.0	24 797	822	846	871	895	920	944	969	993	'018
9	22.5	25 042	066	091	115	139	164	188	212	237	
		285	310	334	358	382	406	431	455	479	
		180	527	551	575	600	624	648	672	696	720
24	23	25 768	792	816	840	864	888	912	935	959	983
2.4	2.3	26 007	031	055	079	102	126	150	174	198	221
4.8	4.6	245	269	293	316	340	364	387	411	435	458
7.2	6.9										
9.6	9.2	482	505	529	553	576	600	623	647	670	694
12.0	11.5	717	741	764	788	811	834	858	881	905	928
14.4	13.8	26 951	975	998	*021	'045	'068	'091	114	138	161
16.8	16.1										
19.2	18.4	27 184	207	231	254	277	300	323	346	370	393
21.6	20.7	416	439	462	485	508	531	554	577	600	623
		646	669	692	715	738	761	784	807	830	852
		190	27 875	898	921	944	967	989	*012	'035	'058
											'081
22	21	91	28 103	126	149	171	194	217	240	262	285
2.2	2.1	92	330	353	375	398	421	443	466	488	511
4.4	4.2	93	556	578	601	623	646	668	691	713	735
6.6	6.3										758
8.8	8.4	94	28 780	803	825	847	870	892	914	937	959
11.0	10.5	95	29 003	026	048	070	092	115	137	159	181
13.2	12.6	96	226	248	270	292	314	336	358	380	403
15.4	14.7										425
17.6	16.8	97	447	469	491	513	535	557	579	601	623
19.8	18.9	98	667	688	710	732	754	776	798	820	842
		99	29 885	907	929	951	973	994	*016	'038	'060
											'081
		200	30 103	125	146	168	190	211	233	255	276
											298

Prop. Parts

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
200	30 103	125	146	168	190	211	233	255	276	298	
01	320	341	363	384	406	428	449	471	492	514	
02	535	557	578	600	621	643	664	685	707	728	
03	750	771	792	814	835	856	878	899	920	942	
04	30 963	984	*006	*027	*048	*069	*091	*112	*133	*154	
05	31 175	197	218	239	260	281	302	323	345	366	
06	387	408	429	450	471	492	513	534	555	576	
07	597	618	639	660	681	702	723	744	765	785	
08	31 806	827	848	869	890	911	931	952	973	994	
09	32 015	035	056	077	098	118	139	160	181	201	
210	222	243	263	284	305	325	346	366	387	408	
11	428	449	469	490	510	531	552	572	593	613	
12	634	654	675	695	715	736	756	777	797	818	
13	32 838	858	879	899	919	940	960	980	*001	*021	
14	33 041	062	082	102	122	143	163	183	203	224	
15	244	264	284	304	325	345	365	385	405	425	
16	445	465	486	506	526	546	566	586	606	626	
17	646	666	686	706	726	746	766	786	806	826	
18	33 846	866	885	905	925	945	965	985	*005	*025	
19	34 044	064	084	104	124	143	163	183	203	223	
220	242	262	282	301	321	341	361	380	400	420	
21	439	459	479	498	518	537	557	577	596	616	
22	635	655	674	694	713	733	753	772	792	811	
23	34 830	850	869	889	908	928	947	967	986	*006	
24	35 025	044	064	083	102	122	141	160	180	199	
25	218	238	257	276	295	315	334	353	372	392	
26	411	430	449	468	488	507	526	545	564	583	
27	603	622	641	660	679	698	717	736	755	774	
28	793	813	832	851	870	889	908	927	946	965	
29	35 984	*003	*021	*040	*059	*078	*097	*116	*135	*154	
230	36 173	192	211	229	248	267	286	*305	324	342	
31	361	380	399	418	436	455	474	493	511	530	
32	549	568	586	605	624	642	661	680	698	717	
33	736	754	773	791	810	829	847	866	884	903	
34	36 922	940	959	977	996	*014	*033	*051	*070	*088	
35	37 107	125	144	162	181	199	218	236	254	273	
36	291	310	328	346	365	383	401	420	438	457	
37	475	493	511	530	548	566	585	603	621	639	
38	658	676	694	712	731	749	767	785	803	822	
39	37 840	858	876	894	912	931	949	967	985	*003	
240	38 021	039	057	075	093	112	*130	148	166	184	
41	202	220	238	256	274	292	310	328	346	364	
42	382	399	417	435	453	471	489	507	525	543	
43	561	578	596	614	632	650	668	686	703	721	
44	739	757	775	792	810	828	846	863	881	899	
45	38 917	934	952	970	987	*005	*023	*041	*058	*076	
46	39 094	111	129	146	164	182	199	217	235	252	
47	270	287	305	322	340	358	375	393	410	428	
48	445	463	480	498	515	533	550	568	585	602	
49	620	637	655	672	690	707	724	742	759	777	
250	39 794	811	829	846	863	881	898	915	933	*950	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

Prop. Parts		N	0	1	2	3	4	5	6	7	8	9
		250	39 794	811	829	846	863	881	898	915	933	950
1 2 3 4 5 6 7 8 9	18	51	39 967	985	*002	*019	*037	*054	*071	*088	*106	*123
	1.8	52	40 140	157	175	192	209	226	243	261	278	295
	3.6	53	312	329	346	364	381	398	415	432	449	466
	5.4	54	483	500	518	535	552	569	586	603	620	637
	7.2	55	654	671	688	705	722	739	756	773	790	807
	9.0	56	824	841	858	875	892	909	926	943	960	976
	10.8											
	12.6	57	40 993	*010	*027	*044	*061	*078	*095	*111	*128	*145
	14.4	58	41 162	179	196	212	229	246	263	280	296	313
	16.2	59	330	347	363	380	397	414	430	447	464	481
		260	497	514	531	547	564	581	597	614	631	647
1 2 3 4 5 6 7 8 9	17	61	664	681	697	714	731	747	764	780	797	814
	1.7	62	830	847	863	880	896	913	929	946	963	979
	3.4	63	41 996	*012	*029	*045	*062	*078	*095	*111	*127	*144
	5.1											
	6.8	64	42 160	177	193	210	226	243	259	275	292	308
	8.5	65	325	341	357	374	390	406	423	439	455	472
	10.2	66	488	504	521	537	553	570	586	602	619	635
	11.9	67	651	667	684	700	716	732	749	765	781	797
	13.6	68	813	830	846	862	878	894	911	927	943	959
	15.3	69	42 975	991	*008	*024	*040	*056	*072	*088	*104	*120
		270	43 136	152	169	185	201	217	233	249	265	281
1 2 3 4 5 6 7 8 9	16	71	297	313	329	345	361	377	393	409	425	441
	1.6	72	457	473	489	505	521	537	553	569	584	600
	3.2	73	616	632	648	664	680	696	712	727	743	759
	4.8	74	775	791	807	823	838	854	870	886	902	917
	6.4	75	43 933	949	965	981	996	*012	*028	*044	*059	*075
	8.0	76	44 091	107	122	138	154	170	185	201	217	232
	9.6											
	11.2	77	248	264	279	295	311	326	342	358	373	389
	12.8	78	404	420	436	451	467	483	498	514	529	545
	14.4	79	560	576	592	607	623	638	654	669	685	700
		280	716	731	747	762	778	793	809	824	840	855
1 2 3 4 5 6 7 8 9	15	81	44 871	886	902	917	932	948	963	979	994	*010
	1.5	82	45 025	040	056	071	086	102	117	133	148	163
	3.0	83	179	194	209	225	240	255	271	286	301	317
	4.5	84	332	347	362	378	393	408	423	439	454	469
	6.0	85	484	500	515	530	545	561	576	591	606	621
	7.5	86	637	652	667	682	697	712	728	743	758	773
	9.0											
	10.5	87	788	803	818	834	849	864	879	894	909	924
	12.0	88	45 939	954	969	984	*000	*015	*030	*045	*060	*075
	13.5	89	46 090	105	120	135	150	165	180	195	210	225
		290	240	255	270	285	300	315	330	345	359	374
1 2 3 4 5 6 7 8 9	14	91	389	404	419	434	449	464	479	494	509	523
	1.4	92	538	553	568	583	598	613	627	642	657	672
	2.8	93	687	702	716	731	746	761	776	790	805	820
	4.2	94	835	850	864	879	894	909	923	938	953	967
	5.6	95	46 982	997	*012	*026	*041	*056	*070	*085	*100	*114
	7.0	96	47 129	144	159	173	188	202	217	232	246	261
	8.4											
	9.8	97	276	290	305	319	334	349	363	378	392	407
	11.2	98	422	436	451	465	480	494	509	524	538	553
	12.6	99	567	582	596	611	625	640	654	669	683	698
		300	47 712	727	741	756	770	784	799	813	828	842
Prop. Parts		N	0	1	2	3	4	5	6	7	8	9

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
300	47 712	727	741	756	770	784	799	813	828	842	
01	47 857	871	885	900	914	929	943	958	972	986	
02	48 001	015	029	044	058	073	087	101	116	130	
03	144	159	173	187	202	216	230	244	259	273	
04	287	302	316	330	344	359	373	387	401	416	
05	430	444	458	473	487	501	515	530	544	558	
06	572	586	601	615	629	643	657	671	686	700	
07	714	728	742	756	770	785	799	813	827	841	
08	855	869	883	897	911	926	940	954	968	982	
09	48 996	*010	*024	*038	*052	*066	*080	*094	*108	*122	
310	49 136	150	164	178	192	206	220	234	248	262	
11	276	290	304	318	332	346	360	374	388	402	
12	415	429	443	457	471	485	499	513	527	541	
13	554	568	582	596	610	624	638	651	665	679	
14	693	707	721	734	748	762	776	790	803	817	
15	831	845	859	872	886	900	914	927	941	955	
16	49 969	982	996	*010	*024	*037	*051	*065	*079	*092	
17	50 106	120	133	147	161	174	188	202	215	229	
18	243	256	270	284	297	311	325	338	352	365	
19	379	393	406	420	433	447	461	474	488	501	
320	515	529	542	556	569	583	596	610	623	637	
21	651	664	678	691	705	718	732	745	759	772	
22	786	799	813	826	840	853	866	880	893	907	
23	50 920	934	947	961	974	987	*001	*014	*028	*041	
24	51 055	068	081	095	108	121	135	148	162	175	
25	188	202	215	228	242	255	268	282	295	308	
26	322	335	348	362	375	388	402	415	428	441	
27	455	468	481	495	508	521	534	548	561	574	
28	587	601	614	627	640	654	667	680	693	706	
29	720	733	746	759	772	786	799	812	825	838	
330	851	865	878	891	904	917	930	943	957	970	
31	51 983	996	*009	*022	*035	*048	*061	*075	*088	*101	
32	52 114	127	140	153	166	179	192	205	218	231	
33	244	257	270	284	297	310	323	336	349	362	
34	375	388	401	414	427	440	453	466	479	492	
35	504	517	530	543	556	569	582	595	608	621	
36	634	647	660	673	686	699	711	724	737	750	
37	763	776	789	802	815	827	840	853	866	879	
38	52 892	905	917	930	943	956	969	982	994	*007	
39	53 020	033	046	058	071	084	097	110	122	135	
340	148	161	173	186	199	212	224	237	250	263	
41	275	288	301	314	326	339	352	364	377	390	
42	403	415	428	441	453	466	479	491	504	517	
43	529	542	555	567	580	593	605	618	631	643	
44	666	668	681	694	706	719	732	744	757	769	
45	782	794	807	820	832	845	857	870	882	895	
46	53 908	920	933	945	958	970	983	995	*008	*020	
47	54 033	045	058	070	083	095	108	120	133	145	
48	158	170	183	195	208	220	233	245	258	270	
49	283	295	307	320	332	345	357	370	382	394	
350	54 407	419	432	444	456	469	481	494	506	518	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

15
1 1.5
2 3.0
3 4.5
4 6.0
5 7.5
6 9.0
7 10.5
8 12.0
9 13.5

14
1 1.4
2 2.8
3 4.2
4 5.6
5 7.0
6 8.4
7 9.8
8 11.2
9 12.6

13
1 1.3
2 2.6
3 3.9
4 5.2
5 6.5
6 7.8
7 9.1
8 10.4
9 11.7

12
1 1.2
2 2.4
3 3.6
4 4.8
5 6.0
6 7.2
7 8.4
8 9.6
9 10.8

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

Prop. Parts		N	0	1	2	3	4	5	6	7	8	9
		350	54 407	419	432	444	456	469	481	494	506	518
		51	531	543	555	568	580	593	605	617	630	642
		52	654	667	679	691	704	716	728	741	753	765
		53	777	790	802	814	827	839	851	864	876	888
13		54	54 900	913	925	937	949	962	974	986	998	*011
1	1.3	55	55 023	035	047	060	072	084	096	108	121	133
2	2.6	56	145	157	169	182	194	206	218	230	242	255
3	3.9											
4	5.2											
5	6.5											
6	7.8	57	267	279	291	303	315	328	340	352	364	376
7	9.1	58	388	400	413	425	437	449	461	473	485	497
8	10.4	59	509	522	534	546	558	570	582	594	606	618
9	11.7											
		360	630	642	654	666	678	691	703	715	727	739
		61	751	763	775	787	799	811	823	835	847	859
		62	871	883	895	907	919	931	943	955	967	979
		63	55 991	*003	*015	*027	*038	*050	*062	*074	*086	*098
13		64	56 110	122	134	146	158	170	182	194	205	217
1	1.2	65	229	241	253	265	277	289	301	312	324	336
2	2.4	66	348	360	372	384	396	407	419	431	443	455
3	3.6											
4	4.8	67	467	478	490	502	514	526	538	549	561	573
5	6.0	68	585	597	608	620	632	644	656	667	679	691
6	7.2	69	703	714	726	738	750	761	773	785	797	808
7	8.4											
8	9.6	370	820	832	844	855	867	879	891	902	914	926
9	10.8											
		71	56 937	949	961	972	984	996	*008	*019	*031	*043
		72	57 054	066	078	089	101	113	124	136	148	159
		73	171	183	194	206	217	229	241	252	264	276
		74	287	299	310	322	334	345	357	368	380	392
		75	403	415	426	438	449	461	473	484	496	507
		76	519	530	542	553	565	576	588	600	611	623
		77	634	646	657	669	680	692	703	715	726	738
		78	749	761	772	784	795	807	818	830	841	852
		79	864	875	887	898	910	921	933	944	955	967
11		380	57 978	990	*001	*013	*024	*035	*047	*058	*070	*081
1	1.1											
2	2.2	81	58 092	104	115	127	138	149	161	172	184	195
3	3.3	82	206	218	229	240	252	263	274	286	297	309
4	4.4	83	320	331	343	354	365	377	388	399	410	422
5	5.5											
6	6.6	84	433	444	456	467	478	490	501	512	524	535
7	7.7	85	546	557	569	580	591	602	614	625	636	647
8	8.8	86	659	670	681	692	704	715	726	737	749	760
9	9.9											
		87	771	782	794	805	816	827	838	850	861	872
		88	883	894	906	917	928	939	950	961	973	984
		89	58 995	*006	*017	*028	*040	*051	*062	*073	*084	*095
10		390	59 106	118	129	140	151	162	173	184	195	207
1	1.0	91	218	229	240	251	262	273	284	295	306	318
2	2.0	92	329	340	351	362	373	384	395	406	417	428
3	3.0	93	439	450	461	472	483	494	506	517	528	539
4	4.0											
5	5.0	94	550	561	572	583	594	605	616	627	638	649
6	6.0	95	660	671	682	693	704	715	726	737	748	759
7	7.0	96	770	780	791	802	813	824	835	846	857	868
8	8.0											
9	9.0											
		97	879	890	901	912	923	934	945	956	966	977
		98	59 988	999	*010	*021	*032	*043	*054	*065	*076	*086
		99	60 097	108	119	130	141	152	163	173	184	195
		400	60 206	217	228	239	249	260	271	282	293	304
Prop. Parts		N	0	1	2	3	4	5	6	7	8	9

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
400	60 206	217	228	239	249	260	271	282	293	304	
01	314	325	336	347	358	369	379	390	401	412	
02	423	433	444	455	466	477	487	498	509	520	
03	531	541	552	563	574	584	595	606	617	627	
04	638	649	660	670	681	692	703	713	724	735	
05	746	756	767	778	788	799	810	821	831	842	
06	853	863	874	885	895	906	917	927	938	949	
07	60 959	970	981	991	*002	*013	*023	*034	*045	*055	11
08	61 066	077	087	098	109	119	130	140	151	162	1 1.1
09	172	183	194	204	215	225	236	247	257	268	2 2.2
410	278	289	300	310	321	331	342	352	363	374	3 3.3
11	384	395	405	416	426	437	448	458	469	479	4 4.4
12	490	500	511	521	532	542	553	563	574	584	5 5.5
13	595	606	616	627	637	648	658	669	679	690	6 6.6
14	700	711	721	731	742	752	763	773	784	794	7 7.7
15	805	815	826	836	847	857	868	878	888	899	8 8.8
16	61 909	920	930	941	951	962	972	982	993	*003	9 9.9
17	62 014	024	034	045	055	066	076	086	097	107	
18	118	128	138	149	159	170	180	190	201	211	
19	221	232	242	252	263	273	284	294	304	315	
420	325	335	346	356	366	377	387	397	408	418	
21	428	439	449	459	469	480	490	500	511	521	10
22	531	542	552	562	572	583	593	603	613	624	1 1.0
23	634	644	655	665	675	685	696	706	716	726	2 2.0
24	737	747	757	767	778	788	798	808	818	829	3 3.0
25	839	849	859	870	880	890	900	910	921	931	4 4.0
26	62 941	951	961	972	982	992	*002	*012	*022	*033	5 5.0
27	63 043	053	063	073	083	094	104	114	124	134	6 6.0
28	144	155	165	175	185	195	205	215	225	236	7 7.0
29	246	256	266	276	286	296	306	317	327	337	8 8.0
430	347	357	367	377	387	397	407	417	428	438	9 9.0
31	448	458	468	478	488	498	508	518	528	538	
32	548	558	568	579	589	599	609	619	629	639	
33	649	659	669	679	689	699	709	719	729	739	
34	749	759	769	779	789	799	809	819	829	839	
35	849	859	869	879	889	899	909	919	929	939	
36	63 949	959	969	979	988	998	*008	*018	*028	*038	
37	64 048	058	068	078	088	098	108	118	128	137	
38	147	157	167	177	187	197	207	217	227	237	
39	246	256	266	276	286	296	306	316	326	335	
440	345	355	365	375	385	395	404	414	424	434	
41	444	454	464	473	483	493	503	513	523	532	
42	542	552	562	572	582	591	601	611	621	631	
43	640	650	660	670	680	689	699	709	719	729	
44	738	748	758	768	777	787	797	807	816	826	
45	836	846	856	865	875	885	895	904	914	924	
46	64 933	943	953	963	972	982	992	*002	*011	*021	
47	65 031	040	050	060	070	079	089	099	108	118	
48	128	137	147	157	167	176	186	196	205	215	
49	225	234	244	254	263	273	283	292	302	312	
450	65 321	331	341	350	360	369	379	389	398	408	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

Prop. Parts		N	0	1	2	3	4	5	6	7	8	9
1 2 3 4 5 6 7 8 9	10	450	65 321	331	341	350	360	369	379	389	398	408
	1.0	51	418	427	437	447	456	466	475	485	495	504
	2.0	52	514	523	533	543	552	562	571	581	591	600
	3.0	53	610	619	629	639	648	658	667	677	686	696
	4.0	54	706	715	725	734	744	753	763	772	782	792
	5.0	55	801	811	820	830	839	849	858	868	877	887
	6.0	56	896	906	916	925	935	944	954	963	973	982
	7.0	57	65 992	*001	*011	*020	*030	*039	*049	*058	*068	*077
	8.0	58	66 087	096	106	115	124	134	143	153	162	172
	9.0	59	181	191	200	210	219	229	238	247	257	266
1 2 3 4 5 6 7 8 9	10	460	276	285	295	304	314	323	332	342	351	361
	1.0	61	370	380	389	398	408	417	427	436	445	455
	2.0	62	464	474	483	492	502	511	521	530	539	549
	3.0	63	558	567	577	586	596	605	614	624	633	642
	4.0	64	652	661	671	680	689	699	708	717	727	736
	5.0	65	745	755	764	773	783	792	801	811	820	829
	6.0	66	839	848	857	867	876	885	894	904	913	922
	7.0	67	66 932	941	950	960	969	978	987	997	*006	*015
	8.0	68	67 025	034	043	052	062	071	080	089	099	108
	9.0	69	117	127	136	145	154	164	173	182	191	201
1 2 3 4 5 6 7 8 9	10	470	210	219	228	237	247	256	265	274	284	293
	1.0	71	302	311	321	330	339	348	357	367	376	385
	2.0	72	394	403	413	422	431	440	449	459	468	477
	3.0	73	486	495	504	514	523	532	541	550	560	569
	4.0	74	578	587	596	605	614	624	633	642	651	660
	5.0	75	669	679	688	697	706	715	724	733	742	752
	6.0	76	761	770	779	788	797	806	815	825	834	843
	7.0	77	852	861	870	879	888	897	906	916	925	934
	8.0	78	67 943	952	961	970	979	988	997	*006	*015	*024
	9.0	79	68 034	043	052	061	070	079	088	097	106	115
1 2 3 4 5 6 7 8 9	10	480	124	133	142	151	160	169	178	187	196	205
	1.0	81	215	224	233	242	251	260	269	278	287	296
	2.0	82	305	314	323	332	341	350	359	368	377	386
	3.0	83	396	404	413	422	431	440	449	458	467	476
	4.0	84	485	494	502	511	520	529	538	547	556	565
	5.0	85	574	583	592	601	610	619	628	637	646	655
	6.0	86	664	673	681	690	699	708	717	726	735	744
	7.0	87	753	762	771	780	789	797	806	815	824	833
	8.0	88	842	851	860	869	878	886	895	904	913	922
	9.0	89	68 931	940	949	958	966	975	984	993	*002	*011
1 2 3 4 5 6 7 8 9	10	490	69 020	028	037	046	055	064	073	082	090	099
	1.0	91	108	117	126	135	144	152	161	170	179	188
	2.0	92	197	205	214	223	232	241	249	258	267	276
	3.0	93	286	294	302	311	320	329	338	346	355	364
	4.0	94	373	381	390	399	408	417	425	434	443	452
	5.0	95	461	469	478	487	496	504	513	522	531	539
	6.0	96	548	557	566	574	583	592	601	609	618	627
	7.0	97	636	644	653	662	671	679	688	697	705	714
	8.0	98	723	732	740	749	758	767	775	784	793	801
	9.0	99	810	819	827	836	845	854	862	871	880	888
500		69 897	906	914	923	932	940	949	958	966	975	
Prop. Parts		N	0	1	2	3	4	5	6	7	8	9

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
500	69 897	906	914	923	932	940	949	958	966	975	
01	69 984	992	*001	*010	*018	*027	*036	*044	*053	*062	
02	70 070	079	088	096	105	114	122	131	140	148	
03	157	165	174	183	191	200	209	217	226	234	
04	243	252	260	269	278	286	295	303	312	321	
05	329	338	346	355	364	372	381	389	398	406	
06	415	424	432	441	449	458	467	475	484	492	
07	501	509	518	526	535	544	552	561	569	578	
08	586	595	603	612	621	629	638	646	655	663	
09	672	680	689	697	706	714	723	731	740	749	
510	757	766	774	783	791	800	808	817	825	834	
11	842	851	859	868	876	885	893	902	910	919	
12	70 927	935	944	952	961	969	978	986	995	*003	
13	71 012	020	029	037	046	054	063	071	079	088	
14	096	105	113	122	130	139	147	155	164	172	
15	181	189	198	206	214	223	231	240	248	257	
16	265	273	282	290	299	307	315	324	332	341	
17	349	357	366	374	383	391	399	408	416	425	
18	433	441	450	458	466	475	483	492	500	508	
19	517	525	533	542	550	559	567	575	584	592	
520	600	609	617	625	634	642	650	659	667	675	
21	684	692	700	709	717	725	734	742	750	759	
22	767	775	784	792	800	809	817	825	834	842	
23	850	858	867	875	883	892	900	908	917	925	
24	71 933	941	950	958	966	975	983	991	999	*008	
25	72 016	024	032	041	049	057	066	074	082	090	
26	099	107	115	123	132	140	148	156	165	173	
27	181	189	198	206	214	222	230	239	247	255	
28	263	272	280	288	296	304	313	321	329	337	
29	346	354	362	370	378	387	395	403	411	419	
530	428	436	444	452	460	469	477	485	493	501	
31	509	518	526	534	542	550	558	567	575	583	
32	591	599	607	616	624	632	640	648	656	665	
33	673	681	689	697	705	713	722	730	738	746	
34	754	762	770	779	787	795	803	811	819	827	
35	835	843	852	860	868	876	884	892	900	908	
36	916	925	933	941	949	957	965	973	981	989	
37	72 997	*006	*014	*022	*030	*038	*046	*054	*062	*070	
38	73 078	086	094	102	111	119	127	135	143	151	
39	159	167	175	183	191	199	207	215	223	231	
540	239	247	255	263	272	280	288	296	304	312	
41	320	328	336	344	352	360	368	376	384	392	
42	400	408	416	424	432	440	448	456	464	472	
43	480	488	496	504	512	520	528	536	544	552	
44	560	568	576	584	592	600	608	616	624	632	
45	640	648	656	664	672	679	687	695	703	711	
46	719	727	735	743	751	759	767	775	783	791	
47	799	807	815	823	830	838	846	854	862	870	
48	878	886	894	902	910	918	926	933	941	949	
49	73 957	965	973	981	989	997	*005	*013	*020	*028	
550	74 036	044	052	060	068	076	084	092	099	107	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

Prop. Parts		N	0	1	2	3	4	5	6	7	8	9
		550	74 036	044	052	060	068	076	084	092	099	107
		51	115	123	131	139	147	155	162	170	178	186
		52	194	202	210	218	225	233	241	249	257	265
		53	273	280	288	296	304	312	320	327	335	343
		54	351	359	367	374	382	390	398	406	414	421
		55	429	437	445	453	461	468	476	484	492	500
		56	507	515	523	531	539	547	554	562	570	578
		57	586	593	601	609	617	624	632	640	648	656
		58	663	671	679	687	695	702	710	718	726	733
		59	741	749	757	764	772	780	788	796	803	811
		560	819	827	834	842	850	858	865	873	881	889
1 2 3 4 5 6 7 8 9	8	61	896	904	912	920	927	935	943	950	958	966
		62	74 974	981	989	997	*005	*012	*020	*028	*035	*043
		63	75 051	059	066	074	082	089	097	105	113	120
		64	128	136	143	151	159	166	174	182	189	197
		65	205	213	220	228	236	243	251	259	266	274
		66	282	289	297	305	312	320	328	335	343	351
		67	358	366	374	381	389	397	404	412	420	427
		68	435	442	450	458	465	473	481	488	496	504
		69	511	519	526	534	542	549	557	565	572	580
		570	587	595	603	610	618	626	633	641	648	656
		71	664	671	679	686	694	702	709	717	724	732
		72	740	747	755	762	770	778	785	793	800	808
		73	815	823	831	838	846	853	861	868	876	884
		74	891	899	906	914	921	929	937	944	952	959
		75	75 967	974	982	989	997	*005	*012	*020	*027	*035
		76	76 042	050	057	065	072	080	087	095	103	110
		77	118	125	133	140	148	155	163	170	178	185
		78	193	200	208	215	223	230	238	245	253	260
		79	268	275	283	290	298	305	313	320	328	335
		580	343	350	358	365	373	380	388	395	403	410
1 2 3 4 5 6 7 8 9	7	81	418	425	433	440	448	455	462	470	477	485
		82	492	500	507	515	522	530	537	545	552	559
		83	567	574	582	589	597	604	612	619	626	634
		84	641	649	656	664	671	678	686	693	701	708
		85	716	723	730	738	745	753	760	768	775	782
		86	790	797	805	812	819	827	834	842	849	856
		87	864	871	879	886	893	901	908	916	923	930
		88	76 938	945	953	960	967	975	982	989	997	*004
		89	77 012	019	026	034	041	048	056	063	070	078
		590	085	093	100	107	115	122	129	137	144	151
		91	159	166	173	181	188	195	203	210	217	225
		92	232	240	247	254	262	269	276	283	291	298
		93	305	313	320	327	335	342	349	357	364	371
		94	379	386	393	401	408	415	422	430	437	444
		95	452	459	466	474	481	488	495	503	510	517
		96	525	532	539	546	554	561	568	576	583	590
		97	597	605	612	619	627	634	641	648	656	663
		98	670	677	685	692	699	706	714	721	728	735
		99	743	750	757	764	772	779	786	793	801	808
		600	77 815	822	830	837	844	851	859	866	873	880
Prop. Parts		N	0	1	2	3	4	5	6	7	8	9

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
600	77 815	822	830	837	844	851	859	866	873	880	
01	887	895	902	909	916	924	931	938	945	952	
02	77 960	967	974	981	988	996	*003	*010	*017	*025	
03	78 032	039	046	053	061	068	075	082	089	097	
04	104	111	118	125	132	140	147	154	161	168	
05	176	183	190	197	204	211	219	226	233	240	
06	247	254	262	269	276	283	290	297	305	312	
07	319	326	333	340	347	355	362	369	376	383	8
08	390	398	405	412	419	426	433	440	447	455	1 0.8
09	462	469	476	483	490	497	504	512	519	526	2 1.6
610	533	540	547	554	561	569	576	583	590	597	3 2.4
11	604	611	618	625	633	640	647	654	661	668	4 3.2
12	675	682	689	696	704	711	718	725	732	739	5 4.0
13	746	753	760	767	774	781	789	796	803	810	6 4.8
14	817	824	831	838	845	852	859	866	873	880	7 5.6
15	888	895	902	909	916	923	930	937	944	951	8 6.4
16	78 958	965	972	979	986	993	*000	*007	*014	*021	9 7.2
17	79 029	036	043	050	057	064	071	078	085	092	
18	099	106	113	120	127	134	141	148	155	162	
19	169	176	183	190	197	204	211	218	225	232	
620	239	246	253	260	267	274	281	288	295	302	
21	309	316	323	330	337	344	351	358	365	372	7
22	379	386	393	400	407	414	421	428	435	442	1 0.7
23	449	456	463	470	477	484	491	498	505	511	2 1.4
24	518	525	532	539	546	553	560	567	574	581	3 2.1
25	588	595	602	609	616	623	630	637	644	650	4 2.8
26	657	664	671	678	685	692	699	706	713	720	5 3.5
27	727	734	741	748	754	761	768	775	782	789	6 4.2
28	796	803	810	817	824	831	837	844	851	858	7 4.9
29	865	872	879	886	893	900	906	913	920	927	8 5.6
630	79 934	941	948	955	962	969	975	982	989	996	9 6.3
31	80 003	010	017	024	030	037	044	051	058	065	
32	072	079	085	092	099	106	113	120	127	134	
33	140	147	154	161	168	175	182	188	195	202	
34	209	216	223	229	236	243	250	257	264	271	
35	277	284	291	298	305	312	318	325	332	339	
36	346	353	359	366	373	380	387	393	400	407	
37	414	421	428	434	441	448	455	462	468	475	
38	482	489	496	502	509	516	523	530	536	543	
39	550	557	564	570	577	584	591	598	604	611	
640	618	625	632	638	645	652	659	665	672	679	
41	686	693	699	706	713	720	726	733	740	747	
42	754	760	767	774	781	787	794	801	808	814	
43	821	828	835	841	848	855	862	868	875	882	
44	889	895	902	909	916	922	929	936	943	949	
45	80 956	963	969	976	983	990	996	*003	*010	*017	
46	81 023	030	037	043	050	057	064	070	077	084	
47	090	097	104	111	117	124	131	137	144	151	
48	158	164	171	178	184	191	198	204	211	218	
49	224	231	238	245	251	258	265	271	278	285	
650	81 291	298	305	311	318	325	331	338	345	351	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

Prop. Parts		N	0	1	2	3	4	5	6	7	8	9
		650	81 291	298	305	311	318	325	331	338	345	351
		51	358	365	371	378	385	391	398	405	411	418
		52	425	431	438	445	451	458	465	471	478	485
		53	491	498	505	511	518	525	531	538	544	551
		54	558	564	571	578	584	591	598	604	611	617
		55	624	631	637	644	651	657	664	671	677	684
		56	690	697	704	710	717	723	730	737	743	750
		57	757	763	770	776	783	790	796	803	809	816
		58	823	829	836	842	849	856	862	869	875	882
		59	889	895	902	908	915	921	928	935	941	948
		600	81 954	961	968	974	981	987	994	*000	*007	*014
		61	82 020	027	033	040	046	053	060	066	073	079
		62	086	092	099	105	112	119	125	132	138	145
		63	151	158	164	171	178	184	191	197	204	210
		64	217	223	230	236	243	249	256	263	269	276
		65	282	289	295	302	308	315	321	328	334	341
		66	347	354	360	367	373	380	387	393	400	406
		67	413	419	426	432	439	445	452	458	465	471
		68	478	484	491	497	504	510	517	523	530	536
		69	543	549	556	562	569	575	582	588	595	601
		670	607	614	620	627	633	640	646	653	659	666
		71	672	679	685	692	698	705	711	718	724	730
		72	737	743	750	756	763	769	776	782	789	795
		73	802	808	814	821	827	834	840	847	853	860
		74	866	872	879	885	892	898	905	911	918	924
		75	930	937	943	950	956	963	969	975	982	988
		76	82 995	*001	*008	*014	*020	*027	*033	*040	*046	*052
		77	83 059	065	072	078	085	091	097	104	110	117
		78	123	129	136	142	149	155	161	168	174	181
		79	187	193	200	206	213	219	225	232	238	245
		680	251	257	264	270	276	283	289	296	302	308
		81	315	321	327	334	340	347	353	359	366	372
		82	378	385	391	398	404	410	417	423	429	436
		83	442	448	455	461	467	474	480	487	493	499
		84	506	512	518	525	531	537	544	550	556	563
		85	569	575	582	588	594	601	607	613	620	626
		86	632	639	645	651	658	664	670	677	683	689
		87	696	702	708	715	721	727	734	740	746	753
		88	759	765	771	778	784	790	797	803	809	816
		89	822	828	835	841	847	853	860	866	872	879
		690	885	891	897	904	910	916	923	929	935	942
		91	83 948	954	960	967	973	979	985	992	998	*004
		92	84 011	017	023	029	036	042	048	055	061	067
		93	073	080	086	092	098	105	111	117	123	130
		94	136	142	148	155	161	167	173	180	186	192
		95	198	205	211	217	223	230	236	242	248	255
		96	261	267	273	280	286	292	298	305	311	317
		97	323	330	336	342	348	354	361	367	373	379
		98	386	392	398	404	410	417	423	429	435	442
		99	448	454	460	466	473	479	485	491	497	504
		700	84 510	516	522	528	535	541	547	553	559	566
Prop. Parts		N	0	1	2	3	4	5	6	7	8	9

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
700	84 510	516	522	528	535	541	547	553	559	566	
01	572	578	584	590	597	603	609	615	621	628	
02	634	640	646	652	658	665	671	677	683	689	
03	696	702	708	714	720	726	733	739	745	751	
04	757	763	770	776	782	788	794	800	807	813	
05	819	825	831	837	844	850	856	862	868	874	
06	880	887	893	899	905	911	917	924	930	936	
07	84 942	948	954	960	967	973	979	985	991	997	7 0.7
08	85 003	009	016	022	028	034	040	046	052	058	1 1.4
09	065	071	077	083	089	095	101	107	114	120	2 2.1
710	126	132	138	144	150	156	163	169	175	181	3 2.8
11	187	193	199	205	211	217	224	230	236	242	4 3.5
12	248	254	260	266	272	278	285	291	297	303	5 4.2
13	309	315	321	327	333	339	345	352	358	364	6 4.9
14	370	376	382	388	394	400	406	412	418	425	7 5.6
15	431	437	443	449	455	461	467	473	479	485	8 6.3
16	491	497	503	509	516	522	528	534	540	546	
17	552	558	564	570	576	582	588	594	600	606	
18	612	618	625	631	637	643	649	655	661	667	
19	673	679	685	691	697	703	709	715	721	727	
720	733	739	745	751	757	763	769	775	781	788	
21	794	800	806	812	818	824	830	836	842	848	
22	854	860	866	872	878	884	890	896	902	908	6 0.6
23	914	920	926	932	938	944	950	956	962	968	1 1.2
24	85 974	980	986	992	998	*004	*010	*016	*022	*028	2 1.8
25	86 034	040	046	052	058	064	070	076	082	088	3 2.4
26	094	100	106	112	118	124	130	136	142	147	4 3.0
27	153	159	165	171	177	183	189	195	201	207	5 3.6
28	213	219	225	231	237	243	249	255	261	267	6 4.2
29	273	279	285	291	297	303	308	314	320	326	7 4.8
730	332	338	344	350	356	362	368	374	380	386	8 5.4
31	392	398	404	410	415	421	427	433	439	445	
32	451	457	463	469	475	481	487	493	499	504	
33	510	516	522	528	534	540	546	552	558	564	
34	570	576	581	587	593	599	605	611	617	623	
35	629	635	641	646	652	658	664	670	676	682	
36	688	694	700	705	711	717	723	729	735	741	
37	747	753	759	764	770	776	782	788	794	800	
38	806	812	817	823	829	835	841	847	853	859	
39	864	870	876	882	888	894	900	906	911	917	
740	923	929	935	941	947	953	958	964	970	976	
41	86 982	988	994	999	*005	*011	*017	*023	*029	*035	5 0.5
42	87 040	046	052	058	064	070	075	081	087	093	1 1.0
43	099	105	111	116	122	128	134	140	146	151	2 1.5
44	157	163	169	175	181	186	192	198	204	210	3 2.0
45	216	221	227	233	239	245	251	256	262	268	4 2.5
46	274	280	286	291	297	303	309	315	320	326	5 3.0
47	332	338	344	349	355	361	367	373	379	384	6 3.5
48	390	396	402	408	413	419	425	431	437	442	7 4.0
49	448	454	460	466	471	477	483	489	495	500	8 4.5
750	87 506	512	518	523	529	535	541	547	552	558	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

Prop. Parts		N	0	1	2	3	4	5	6	7	8	9
<div> <div>6</div> <div>0.6</div> <div>1.2</div> <div>1.8</div> <div>2.4</div> <div>3.0</div> <div>3.6</div> <div>4.2</div> <div>4.8</div> <div>5.4</div> </div>		750	87 506	512	518	523	529	535	541	547	552	558
		51	564	570	576	581	587	593	599	604	610	616
		52	622	628	633	639	645	651	656	662	668	674
		53	679	685	691	697	703	708	714	720	726	731
		54	737	743	749	754	760	766	772	777	783	789
		55	795	800	806	812	818	823	829	835	841	846
		56	852	858	864	869	875	881	887	892	898	904
		57	910	915	921	927	933	938	944	950	955	961
		58	87 967	973	978	984	990	996	*001	*007	*013	*018
		59	88 024	030	036	041	047	053	058	064	070	076
		760	081	087	093	098	104	110	116	121	127	133
		61	138	144	150	156	161	167	173	178	184	190
		62	195	201	207	213	218	224	230	235	241	247
		63	252	258	264	270	275	281	287	292	298	304
		64	309	315	321	326	332	338	343	349	355	360
		65	366	372	377	383	389	395	400	406	412	417
		66	423	429	434	440	446	451	457	463	468	474
		67	480	485	491	497	502	508	513	519	525	530
		68	536	542	547	553	559	564	570	576	581	587
		69	593	598	604	610	615	621	627	632	638	643
		770	649	655	660	666	672	677	683	689	694	700
		71	705	711	717	722	728	734	739	745	750	756
		72	762	767	773	779	784	790	795	801	807	812
		73	818	824	829	835	840	846	852	857	863	868
		74	874	880	885	891	897	902	908	913	919	925
		75	930	936	941	947	953	958	964	969	975	981
		76	88 986	992	997	*003	*009	*014	*020	*025	*031	*037
		77	89 042	048	053	059	064	070	076	081	087	092
		78	098	104	109	115	120	126	131	137	143	148
		79	154	159	165	170	176	182	187	193	198	204
<div> <div>5</div> <div>0.5</div> <div>1.0</div> <div>1.5</div> <div>2.0</div> <div>2.5</div> <div>3.0</div> <div>3.5</div> <div>4.0</div> <div>4.5</div> </div>		780	209	215	221	226	232	237	243	248	254	260
		81	265	271	276	282	287	293	298	304	310	315
		82	321	326	332	337	343	348	354	360	365	371
		83	376	382	387	393	398	404	409	415	421	426
		84	432	437	443	448	454	459	465	470	476	481
		85	487	492	498	504	509	515	520	526	531	537
		86	542	548	553	559	564	570	575	581	586	592
		87	597	603	609	614	620	625	631	636	642	647
		88	653	658	664	669	675	680	686	691	697	702
		89	708	713	719	724	730	735	741	746	752	757
		790	763	768	774	779	785	790	796	801	807	812
		91	818	823	829	834	840	845	851	856	862	867
		92	873	878	883	889	894	900	905	911	916	922
		93	927	933	938	944	949	955	960	966	971	977
		94	89 982	988	993	998	*004	*009	*015	*020	*026	*031
		95	90 037	042	048	053	059	064	069	075	080	086
		96	091	097	102	108	113	119	124	129	135	140
		97	146	151	157	162	168	173	179	184	189	195
		98	200	206	211	217	222	227	233	238	244	249
		99	255	260	266	271	276	282	287	293	298	304
		800	90 309	314	320	325	331	336	342	347	352	358
Prop. Parts		N	0	1	2	3	4	5	6	7	8	9

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
800	90 309	314	320	325	331	336	342	347	352	358	
01	363	369	374	380	385	390	396	401	407	412	
02	417	423	428	434	439	445	450	455	461	466	
03	472	477	482	488	493	499	504	509	515	520	
04	526	531	536	542	547	553	558	563	569	574	
05	580	585	590	596	601	607	612	617	623	628	
06	634	639	644	650	655	660	666	671	677	682	
07	687	693	698	703	709	714	720	725	730	736	
08	741	747	752	757	763	768	773	779	784	789	
09	795	800	806	811	816	822	827	832	838	843	
810	849	854	859	865	870	875	881	886	891	897	
11	902	907	913	918	924	929	934	940	945	950	
12	90 956	961	966	972	977	982	988	993	998	*004	
13	91 009	014	020	025	030	036	041	046	052	057	
14	062	068	073	078	084	089	094	100	105	110	
15	116	121	126	132	137	142	148	153	158	164	
16	169	174	180	185	190	196	201	206	212	217	
17	222	228	233	238	243	249	254	259	265	270	
18	275	281	286	291	297	302	307	312	318	323	
19	328	334	339	344	350	355	360	365	371	376	
820	381	387	392	397	403	408	413	418	424	429	
21	434	440	445	450	455	461	466	471	477	482	
22	487	492	498	503	508	514	519	524	529	535	
23	540	545	551	556	561	566	572	577	582	587	
24	593	598	603	609	614	619	624	630	635	640	
25	645	651	656	661	666	672	677	682	687	693	
26	698	703	709	714	719	724	730	735	740	745	
27	751	756	761	766	772	777	782	787	793	798	
28	803	808	814	819	824	829	834	840	845	850	
29	855	861	866	871	876	882	887	892	897	903	
830	908	913	918	924	929	934	939	944	950	955	
31	91 960	965	971	976	981	986	991	997	*002	*007	
32	92 012	018	023	028	033	038	044	049	054	059	
33	065	070	075	080	085	091	096	101	106	111	
34	117	122	127	132	137	143	148	153	158	163	
35	169	174	179	184	189	195	200	205	210	215	
36	221	226	231	236	241	247	252	257	262	267	
37	273	278	283	288	293	298	304	309	314	319	
38	324	330	335	340	345	350	355	361	366	371	
39	376	381	387	392	397	402	407	412	418	423	
840	428	433	438	443	449	454	459	464	469	474	
41	480	485	490	495	500	505	511	516	521	526	
42	531	536	542	547	552	557	562	567	572	578	
43	583	588	593	598	603	609	614	619	624	629	
44	634	639	645	650	655	660	665	670	675	681	
45	686	691	696	701	706	711	716	722	727	732	
46	737	742	747	752	758	763	768	773	778	783	
47	788	793	799	804	809	814	819	824	829	834	
48	840	845	850	855	860	865	870	875	881	886	
49	891	896	901	906	911	916	921	927	932	937	
850	92 942	947	952	957	962	967	973	978	983	988	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

1 0.6
 2 1.2
 3 1.8
 4 2.4
 5 3.0
 6 3.6
 7 4.2
 8 4.8
 9 5.4

1 0.5
 2 1.0
 3 1.5
 4 2.0
 5 2.5
 6 3.0
 7 3.5
 8 4.0
 9 4.5

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

Prop. Parts		N	0	1	2	3	4	5	6	7	8	9
<div> <div>6</div> <div>0.6</div> <div>1.2</div> <div>1.8</div> <div>2.4</div> <div>3.0</div> <div>3.6</div> <div>4.2</div> <div>4.8</div> <div>5.4</div> </div>		850	92 942	947	952	957	962	967	973	978	983	988
		51	92 993	998	*003	*008	*013	*018	*024	*029	*034	*039
		52	93 044	049	054	059	064	069	075	080	085	090
		53	095	100	105	110	115	120	125	131	136	141
		54	146	151	156	161	166	171	176	181	186	192
		55	197	202	207	212	217	222	227	232	237	242
		56	247	252	258	263	268	273	278	283	288	293
		57	298	303	308	313	318	323	328	334	339	344
		58	349	354	359	364	369	374	379	384	389	394
		59	399	404	409	414	420	425	430	435	440	445
<div> <div>5</div> <div>0.5</div> <div>1.0</div> <div>1.5</div> <div>2.0</div> <div>2.5</div> <div>3.0</div> <div>3.5</div> <div>4.0</div> <div>4.5</div> </div>		860	450	455	460	465	470	475	480	485	490	495
		61	500	505	510	515	520	526	531	536	541	546
		62	551	556	561	566	571	576	581	586	591	596
		63	601	606	611	616	621	626	631	636	641	646
		64	651	656	661	666	671	676	682	687	692	697
		65	702	707	712	717	722	727	732	737	742	747
		66	752	757	762	767	772	777	782	787	792	797
		67	802	807	812	817	822	827	832	837	842	847
		68	852	857	862	867	872	877	882	887	892	897
		69	902	907	912	917	922	927	932	937	942	947
<div> <div>4</div> <div>0.4</div> <div>0.8</div> <div>1.2</div> <div>1.6</div> <div>2.0</div> <div>2.4</div> <div>2.8</div> <div>3.2</div> <div>3.6</div> </div>		870	93 952	957	962	967	972	977	982	987	992	997
		71	94 002	007	012	017	022	027	032	037	042	047
		72	052	057	062	067	072	077	082	086	091	096
		73	101	106	111	116	121	126	131	136	141	146
		74	151	156	161	166	171	176	181	186	191	196
		75	201	206	211	216	221	226	231	236	240	245
		76	250	255	260	265	270	275	280	285	290	295
		77	300	305	310	315	320	325	330	335	340	345
		78	349	354	359	364	369	374	379	384	389	394
		79	399	404	409	414	419	424	429	433	438	443
<div> <div>3</div> <div>0.3</div> <div>0.7</div> <div>1.1</div> <div>1.5</div> <div>1.9</div> <div>2.3</div> <div>2.7</div> <div>3.1</div> <div>3.5</div> </div>		880	448	453	458	463	468	473	478	483	488	493
		81	498	503	507	512	517	522	527	532	537	542
		82	547	552	557	562	567	571	576	581	586	591
		83	596	601	606	611	616	621	626	630	635	640
		84	645	650	655	660	665	670	675	680	685	689
		85	694	699	704	709	714	719	724	729	734	738
		86	743	748	753	758	763	768	773	778	783	787
		87	792	797	802	807	812	817	822	827	832	836
		88	841	846	851	856	861	866	871	876	880	885
		89	890	895	900	905	910	915	919	924	929	934
<div> <div>2</div> <div>0.2</div> <div>0.6</div> <div>1.0</div> <div>1.4</div> <div>1.8</div> <div>2.2</div> <div>2.6</div> <div>3.0</div> <div>3.4</div> </div>		890	939	944	949	954	959	963	968	973	978	983
		91	94 988	993	998	*002	*007	*012	*017	*022	*027	*032
		92	95 036	041	046	051	056	061	066	071	075	080
		93	085	090	095	100	105	109	114	119	124	129
		94	134	139	143	148	153	158	163	168	173	177
		95	182	187	192	197	202	207	211	216	221	226
		96	231	236	240	245	250	255	260	265	270	274
		97	279	284	289	294	299	303	308	313	318	323
		98	328	332	337	342	347	352	357	361	366	371
		99	376	381	386	390	395	400	405	410	415	419
Prop. Parts		N	0	1	2	3	4	5	6	7	8	9
		900	95 424	429	434	439	444	448	453	458	463	468

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

N	0	1	2	3	4	5	6	7	8	9	Prop. Parts
900	96 424	429	434	439	444	448	453	458	463	468	<div> <div>5</div> <div>0.5</div> <div>1.0</div> <div>1.5</div> <div>2.0</div> <div>2.5</div> <div>3.0</div> <div>3.5</div> <div>4.0</div> <div>4.5</div> </div>
01	472	477	482	487	492	497	501	506	511	516	
02	521	525	530	535	540	545	550	554	559	564	
03	569	574	578	583	588	593	598	602	607	612	
04	617	622	626	631	636	641	646	650	655	660	
05	665	670	674	679	684	689	694	698	703	708	
06	713	718	722	727	732	737	742	746	751	756	
07	761	766	770	775	780	785	789	794	799	804	
08	809	813	818	823	828	832	837	842	847	852	
09	856	861	866	871	875	880	885	890	895	899	
910	904	909	914	918	923	928	933	938	942	947	
11	952	957	961	966	971	976	980	985	990	995	
12	95 999	*004	*009	*014	*019	*023	*028	*033	*038	*042	
13	96 047	052	057	061	066	071	076	080	085	090	
14	095	099	104	109	114	118	123	128	133	137	
15	142	147	152	156	161	166	171	175	180	185	
16	190	194	199	204	209	213	218	223	227	232	
17	237	242	246	251	256	261	265	270	275	280	
18	284	289	294	298	303	308	313	317	322	327	
19	332	336	341	346	350	355	360	365	369	374	
920	379	384	388	393	398	402	407	412	417	421	
21	426	431	435	440	445	450	454	459	464	468	<div> <div>4</div> <div>0.4</div> <div>0.8</div> <div>1.2</div> <div>1.6</div> <div>2.0</div> <div>2.4</div> <div>2.8</div> <div>3.2</div> <div>3.6</div> </div>
22	473	478	483	487	492	497	501	506	511	515	
23	520	525	530	534	539	544	548	553	558	562	
24	567	572	577	581	586	591	595	600	605	609	
25	614	619	624	628	633	638	642	647	652	656	
26	661	666	670	675	680	685	689	694	699	703	
27	708	713	717	722	727	731	736	741	745	750	
28	755	759	764	769	774	778	783	788	792	797	
29	802	806	811	816	820	825	830	834	839	844	
930	848	853	858	862	867	872	876	881	886	890	
31	895	900	904	909	914	918	923	928	932	937	
32	942	946	951	956	960	965	970	974	979	984	
33	96 988	993	997	*002	*007	*011	*016	*021	*025	*030	
34	97 035	039	044	049	053	058	063	067	072	077	
35	081	086	090	095	100	104	109	114	118	123	
36	128	132	137	142	146	151	155	160	165	169	
37	174	179	183	188	192	197	202	206	211	216	
38	220	225	230	234	239	243	248	253	257	262	
39	267	271	276	280	285	290	294	299	304	308	
940	313	317	322	327	331	336	340	345	350	354	
41	359	364	368	373	377	382	387	391	396	400	<div> <div>3</div> <div>0.3</div> <div>0.7</div> <div>1.1</div> <div>1.5</div> <div>1.9</div> <div>2.3</div> <div>2.7</div> <div>3.1</div> <div>3.5</div> </div>
42	405	410	414	419	424	428	433	437	442	447	
43	451	456	460	465	470	474	479	483	488	493	
44	497	502	506	511	516	520	525	529	534	539	
45	543	548	552	557	562	566	571	575	580	585	
46	589	594	598	603	607	612	617	621	626	630	
47	635	640	644	649	653	658	663	667	672	676	
48	681	685	690	695	699	704	708	713	717	722	
49	727	731	736	740	745	749	754	759	763	768	
950	772	777	782	786	791	795	800	804	809	813	
N	0	1	2	3	4	5	6	7	8	9	Prop. Parts

TABLE II. COMMON LOGARITHMS OF NUMBERS TO FIVE DECIMAL PLACES

Prop. Parts		N	0	1	2	3	4	5	6	7	8	9
<div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> <div>5</div> <div>6</div> <div>7</div> <div>8</div> <div>9</div> </div> <div> <div>0.5</div> <div>1.0</div> <div>1.5</div> <div>2.0</div> <div>2.5</div> <div>3.0</div> <div>3.5</div> <div>4.0</div> <div>4.5</div> </div>		950	97 772	777	782	786	791	795	800	804	809	813
		51	818	823	827	832	836	841	845	850	855	859
		52	864	868	873	877	882	886	891	896	900	905
		53	909	914	918	923	928	932	937	941	946	950
		54	97 955	959	964	968	973	978	982	987	991	996
		55	98 000	005	009	014	019	023	028	032	037	041
		56	046	050	055	059	064	068	073	078	082	087
		57	091	096	100	105	109	114	118	123	127	132
		58	137	141	146	150	155	159	164	168	173	177
		59	182	186	191	195	200	204	209	214	218	223
		960	227	232	236	241	245	250	254	259	263	268
		61	272	277	281	286	290	295	299	304	308	313
		62	318	322	327	331	336	340	345	349	354	358
		63	363	367	372	376	381	385	390	394	399	403
		64	408	412	417	421	426	430	435	439	444	448
		65	453	457	462	466	471	475	480	484	489	493
		66	498	502	507	511	516	520	525	529	534	538
		67	543	547	552	556	561	565	570	574	579	583
		68	588	592	597	601	605	610	614	619	623	628
		69	632	637	641	646	650	655	659	664	668	673
		970	677	682	686	691	695	700	704	709	713	717
		71	722	726	731	735	740	744	749	753	758	762
		72	767	771	776	780	784	789	793	798	802	807
		73	811	816	820	825	829	834	838	843	847	851
		74	856	860	865	869	874	878	883	887	892	896
		75	900	905	909	914	918	923	927	932	936	941
		76	945	949	954	958	963	967	972	976	981	985
		77	98 989	994	998	*003	*007	*012	*016	*021	*025	*029
		78	99 034	038	043	047	052	056	061	065	069	074
		79	078	083	087	092	096	100	105	109	114	118
		980	123	127	131	136	140	145	149	154	158	162
<div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> <div>5</div> <div>6</div> <div>7</div> <div>8</div> <div>9</div> </div> <div> <div>0.4</div> <div>0.8</div> <div>1.2</div> <div>1.6</div> <div>2.0</div> <div>2.4</div> <div>2.8</div> <div>3.2</div> <div>3.6</div> </div>		81	167	171	176	180	185	189	193	198	202	207
		82	211	216	220	224	229	233	238	242	247	251
		83	255	260	264	269	273	277	282	286	291	295
		84	300	304	308	313	317	322	326	330	335	339
		85	344	348	352	357	361	366	370	374	379	383
		86	388	392	396	401	405	410	414	419	423	427
		87	432	436	441	445	449	454	458	463	467	471
		88	476	480	484	489	493	498	502	506	511	515
		89	520	524	528	533	537	542	546	550	555	559
		990	564	568	572	577	581	585	590	594	599	603
		91	607	612	616	621	625	629	634	638	642	647
		92	651	656	660	664	669	673	677	682	686	691
		93	695	699	704	708	712	717	721	726	730	734
		94	739	743	747	752	756	760	765	769	774	778
		95	782	787	791	795	800	804	808	813	817	822
		96	826	830	835	839	843	848	852	856	861	865
		97	870	874	878	883	887	891	896	900	904	909
		98	913	917	922	926	930	935	939	944	948	952
		99	99 957	961	965	970	974	978	983	987	991	996
		1000	00 000	004	009	013	017	022	026	030	035	039
Prop. Parts		N	0	1	2	3	4	5	6	7	8	9

INDEX

- Arithmetic mean, 33**
 - short methods of computing, 39
 - of sub-sets, 44, 193
- Array, 193**
- Asymmetry, *see* skewness**
- Averages, Chapter III**
 - discussion of different, 51, 52-58
- Average deviation, *see* mean deviation**
- Burr, I. W., 111 ft. nt.**
- Charlier check, 66, 87**
- Charts, 24**
 - ratio, 157
- Classification of data, 9-15**
- Class**
 - boundary, 15
 - interval, 11
 - limits, 15
 - marks, 11
 - mid-value of, 11
- Coefficient**
 - of alienation, 185
 - of correlation, Chapter VII
 - of variation, 90
- Collateral reading, 5**
- Combination of sets, 99**
- Compound interest law, 156**
- Computing machines, 4, 71**
- Constant, 7**
- Correlation**
 - and regression, 178
 - coefficient, Chapter VIII
 - rank, 222
 - ratio, 212
 - relation to common causes, 225
 - interpretation of, 225
- intraclass, 232**
- surface, 208**
- table, 189**
- Cumulative frequencies, 16, 27, 132**
- Curve of error, *see* normal curve**
- Curve fitting, Chapter VII, 124**
- Curves of growth, 53, 152, 164, 166**
- Deviation, 36**
 - mean or average, 84
 - root-mean-square, 87 ft. nt., 99
- Dispersion, *see* measures of, relative 90**
- Dwyer, P. S., 176**
- Estimate, standard error of, 179**
- Frequency**
 - curves, 25, 112
 - distributions, Chapter I
 - graphical representation of, Chapter II
 - polygon, 24
- Function. definition, 22**
 - exponential, 152
 - frequency, 112
 - linear, 137
 - parabolic, 162
 - quadratic, 138
- Geometric mean, 52**
- Gompertz curve, 164**
- Graduation by means of normal curve, 128**
- Graphical representation, Chapters II, VII**
- Harmonic mean, 55**
- Histogram, 25**
- Hotelling, H., 167 ft. nt.**

- Huntington, E. V., 167
Kendall, M. G., 51 ft. nt.
 Kurtosis, 73, 109
Least-squares method, 144
 Logarithmic paper, 161
 Logistic curve, 166
Makeham's law, 167
Mean
 arithmetic, 33
 geometric, 52
 harmonic, 55
 of means, 43
 Mean deviation, 84
 Measures of dispersion, Chapter V
 mean deviation, 84
 quartiles, 82
 semi-interquartile range, 82
 standard deviation, 86
 Median, 47
 Mode, 47
 Moment of a distribution, Chapter IV
 method of, 141
Normal curve, Chapter VI
 explanation of tables of, 116
 fitted to observed data, 124
 properties of, 118
 standard form of, 115
 Normal equations, 145
Ogive, 27
Parabola, fitting a, 162
 Parameter, 115, 124, 141, 153
 Percentiles, 84
 Probability, 131
 Probability paper, 132
Quartiles, 82
 of normal curve, 119
Range, 16
 Ratio charts, 157
 Reed-Pearl curve, 166
 Regression
 coefficients, 178
 linear, 177
 non-linear, 212
 testing linearity of, 217
 Residuals, 144
 Rietz, H. L., 114 ft. nt.
Scatter diagram, 171
 Semi-logarithmic paper, 157
 Sheppard's corrections, 78, 88
 Shewhart, W. A., 75
 Skewness, 73, 109
 Snedecor, G. W., 178
 Standard units, 69
 Statistic, 124
 Standard deviation, 68
 of combination of sets, 99
 of grouped data, 86
 of ungrouped data, 93
 Straight line, 137
 fitting to data, 140
 Symmetry, 73, 109
Tables
 areas under normal curve, Appendix
 logarithms of numbers, Appendix
 ordinates of normal curve Appendix
 Tabulation, 9
 Time series, 150
 Translation of axes, 36
 Trend, 150, 162
Variability, *see* dispersion
 Variable, 7
Variance, 87
 Variates, 7
Walker, Helen M., 199
 Weighted mean, 33
 Wilkens, J. E., 74 ft. nt.
 Wilson, E. B., 226

